Ecological Applications: Article

What can we learn from 100,000 freshwater forecasts? A synthesis from the NEON Ecological Forecasting Challenge

Freya Olsson[1,2*], Cayelan C. Carey[1,2], Carl Boettiger[3], Gregory Harrison[2], Robert Ladwig[4,5], Marcus F. Lapeyrolerie[3], Abigail S. L. Lewis[1], Mary E. Lofton[1,2], Felipe Montealegre-Mora[3], Joseph S. Rabaey[6], Caleb J. Robbins[7,8], Xiao Yang[9], R. Quinn Thomas[1,2,10]

[1]Department of Biological Sciences, Virginia Tech, Blacksburg, Virginia 24061, USA

[2]Center for Ecosystem Forecasting, Virginia Tech, Blacksburg, Virginia 24061, USA

[3]Department of Environmental Science, Policy, and Management, University of California Berkeley, Berkeley, California 94720, USA

[4]Department of Ecoscience, Aarhus University, Aarhus, 8000, Denmark

[5]Center for Limnology, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA

[6]Large Lakes Observatory, University of Minnesota, Duluth, Minnesota 55812, USA

[7]Institute of Arctic Biology, University of Alaska Fairbanks, Fairbanks, Alaska 99775, USA

[8]Center for Reservoir and Aquatic Systems Research, Baylor University, Waco, Texas 76706, USA

[9]Department of Earth Sciences, Southern Methodist University, Dallas, Texas 75275, USA

[10]Department of Forest Resources and Environmental Conservation, Virginia Tech, Blacksburg, Virginia 24061, USA

*Corresponding and contact author (freyao@vt.edu)

Open Research statement

All data are published and publicly available in the Zenodo repository (Olsson et al. 2024a) and cited in this manuscript. The code used to generate figures and results in this manuscript are also archived in the Zenodo repository (Olsson et al. 2024b) and are cited within the manuscript.

Key words: ecological forecasting; forecasting challenge; freshwater; near-term forecast; NEON; uncertainty; water quality

Abstract

Near-term, iterative ecological forecasts can be used to help understand and proactively manage ecosystems. To date, more forecasts have been developed for aquatic ecosystems than other ecosystems worldwide, likely motivated by the pressing need to conserve these essential and threatened ecosystems. Forecasters have implemented many different modelling approaches to forecast freshwater variables, which have demonstrated promise at individual sites. However, a comprehensive analysis of the performance of varying forecast models across multiple sites is needed to understand broader controls on forecast performance. Forecasting challenges (i.e., community-scale efforts to generate forecasts while also developing shared software, training materials, and best practices) present a useful platform for bridging this gap to evaluate how a range of modelling methods perform across axes of space, time, and ecological systems. Here, we analysed forecasts from the aquatics theme of the National Ecological Observatory Network (NEON) Forecasting Challenge hosted by the Ecological Forecasting Initiative. Over 100,000 probabilistic forecasts of water temperature and dissolved oxygen concentration for 1-30 days ahead across seven NEON-monitored lakes were submitted in 2023. We assessed how forecast performance varied among models with different structures, covariates, and sources of uncertainty relative to baseline null models. More models outperformed the baseline models in forecasting water temperature (ten models) than dissolved oxygen (six). These top-performing models came from a range of classes and structures. For water temperature, we found that process-based models and models that included air temperature as a covariate generally exhibited the highest forecast performance across all sites, and that the most skillful forecasts often accounted for more sources of uncertainty than the lower-performing models. The most skillful forecasts were observed at sites where observations were most divergent from historical

conditions (resulting in poor baseline model performance). Overall, the NEON Forecasting Challenge provides an exciting opportunity for a model inter-comparison to learn about the relative strengths of a diverse suite of models and advance our understanding of freshwater ecosystem predictability.

1 Introduction

Ecological forecasting is a growing field that leverages predictions of future ecological states to help understand and manage ecosystems (Tulloch et al., 2020; Lewis et al., 2023; Dietze et al., 2018). Here, we define forecasts as predictions of future conditions with specified uncertainty (Lewis et al. 2022). As environmental conditions increasingly change in response to altered climate and land use (IPCC, 2023), ecological forecasts have considerable potential for improving management to support ecosystem services now and in the future (Bradford et al., 2018, Dietze et al., 2018). Moreover, forecasting future conditions that have yet to occur inherently requires out-of-sample implementation of models, which can lead to insights into optimal modelling approaches (Lewis et al., 2023).

In freshwater ecosystems, rapid environmental change has led to conditions that are both more variable and outside of historically observed states, motivating a particular need for near-term, iterative ecological forecasts (e.g., Carey, 2023, Richardson et al., 2024; Siam & Eltahir 2017). Near-term (i.e., sub-daily to decadal) forecasts allow researchers to evaluate models within management-relevant timescales (Dietze et al., 2018), and iteratively updating and evaluating forecasts enables rapid improvement in forecast performance by integrating observational data and updating parameters (Dietze et al., 2018, Loescher et al., 2017). These near-term iterative ecological forecasts will help protect critical provisioning, regulating,

supporting, and cultural services (Sterner et al., 2020; Dodds et al., 2013; Lofton et al. 2023) that these highly threatened systems provide (Carrizo et al., 2017; Dudgeon, et al., 2006; Reid et al., 2019), thereby improving management and mitigation (e.g. Huang et al., 2011; Carey et al., 2022; Zwart et al., 2023).

Although the number of near-term, iterative water quality forecasts of freshwater ecosystems is growing (Lofton et al., 2023), challenges remain in producing reliable and accurate predictions of changes in these environments. To date, researchers have implemented many classes of models to forecast freshwater variables (reviewed by Lofton et al. 2023), including process-based (PB) models (Baracchini et al, 2020; Clayer et al., 2023, Thomas et al., 2020, Page et al., 2018), machine learning (ML) models (Di Nunno 2023; Cheng et al., 2020; Read et al., 2019; Zwart et al., 2023), statistical models (Woelmer et al. 2022, McClure et al. 2021; Caissie et al., 2017), and multi-model and hybrid approaches (Olsson et al., 2024c, Saber 2020; Qu et al., 2017). In addition, forecasts have been generated using a range of model covariates (i.e., driver variables). In many cases, weather forecasts are used as covariates because meteorology is a key driver of many ecosystem processes in freshwater ecosystems (Hipsey et al 2019; Livingstone & Padisák, 2007, Rousso et al., 2020). Additionally, some models include autoregressive terms as covariates (e.g., ARIMA models). While forecasting methods have demonstrated promise at individual freshwater sites or a handful of sites (e.g., Barrachini et al., 2020; Thomas et al., 2020; Zwart et al., 2023; Oullex-Proulx et al., 2017, Page et al., 2018; Chen et al., 2024), to date there has yet to be a comprehensive analysis of the performance of forecasting models across a large range of model classes and model covariates across multiple sites.

Forecasting challenges present a useful platform for bridging this gap and learning about how a range of modelling methods perform across axes of space, time, and ecological systems

(Thomas et al., 2023; Humphries et al., 2018). Forecasting challenges typically entail an open call to the research community with a 'challenge' to forecast a specific variable, standardised requirements, and formal evaluation of out-of-sample time steps. Some challenges have aimed to identify a "winner" or best approach, while others have focused more on community and knowledge building (Thomas et al., 2023; Makridakis et al., 2020; Humphries et al., 2018). By bringing together individuals and teams from broad backgrounds, challenges provide opportunities for innovation and community-building, and the development of community cyberinfrastructure can accelerate discipline-wide progress (Fer et al., 2021). Altogether, this collaborative effort can facilitate the development of new methods, standardisation of forecasting targets and formats, and tools and templates that expand the training and education to improve accessibility of forecasting (Thomas et al., 2023). While forecasting challenges are common in the fields of finance, business, demography (Makridakis et al., 2020; Bojer & Meldgaard 2021), and epidemiology (Johansson et al., 2019; Viboud et al., 2018; Biggerstaff et al., 2018), few have existed in ecology until recently (e.g., Humphries et al., 2018, Wheeler et al., 2024), providing new opportunities for advancing the discipline. For example, previous efforts to compare outcomes among ecological forecasting methods have been hindered by differences in evaluation metrics, sites, and variables being forecasted (e.g., Ruosso et al., 2020), which can be addressed by a standardised forecasting challenge framework.

The National Ecological Observatory Network (NEON) Forecasting Challenge (hereafter NEON Challenge), hosted by the Ecological Forecasting Initiative (EFI) Research Coordination Network, was designed to initiate these advances in ecological forecasting. The NEON Challenge is "an open platform for the ecological and data science communities to forecast NEON data before they are collected" (Thomas et al., 2023). The challenge aims to galvanise the

forecasting community around a common framework, with the goals of: improving forecasting tools (e.g., Dietze et al., 2023), learning about ecological predictability (e.g., Wheeler et al., 2024), and advancing training (e.g., Willson et al., 2023).

The NEON Challenge provides a unique case study for examining the performance of freshwater forecasts across space, time, and ecological systems. Ecological time-series present specific complexities compared to previous forecasting challenges given the variability in ecological data collection, irregularities in data resolution, and the inherent variability of the observations (Farley et al., 2018; Michener & Jones, 2012). Moreover, unlike previous forecasting challenges, the NEON Challenge is on-going and accepts submissions of as-yet-unmeasured conditions on a rolling basis with scoring occurring continuously as new data are collected and made available in near real-time (Thomas et al., 2023). In the aquatics lake theme of the NEON Challenge, participants were invited to submit 1 to 30 day-ahead probabilistic forecasts of daily surface mean water temperature (hereafter, Tw) and dissolved oxygen concentration (DO) of seven NEON lake sites, with new forecasts accepted daily (Thomas et al. 2023). Due to issues relating to data quality, submitted forecasts of chlorophyll *a* were omitted from our analysis. Forecasts were solicited across a range of sites, dates, and variables to understand how skill varies across these three axes. Forecasts could be generated using any method but had to include an estimate of uncertainty.

The inclusion of, and emphasis on, uncertainty was a novel component of the NEON Challenge, as uncertainty has been rarely included in previous forecasting challenges. Meaningful representations of uncertainty are critical to forecast interpretation and comparison, but uncertainty quantification is still not ubiquitous across ecological forecasts (reviewed by Lewis et al., 2022), and freshwater forecasts in particular. In a review of freshwater forecasts by

Lofton et al. (2023), only 16 out 61 near-term (sub-daily to decadal) forecasts of water quality variables included an estimate of the uncertainty associated with a prediction. Uncertainty can arise from a variety of sources: model process, model parameters, model initial conditions, model drivers, and observations (Table 1). The relative importance of each source is often dependent on the ecosystem process or state being forecasted and the forecast horizon (Thomas et al., 2020; Lofton et al., 2022; Ouellet-Proulx et al., 2017).

We were specifically focused on uncertainty in our analysis because forecasts that have good accuracy and good precision (e.g., accurate uncertainty) have been shown to improve decision-making outcomes (Mylne 2009; Nadav-Greenberg et al., 2009; Ramos et al., 2013). NEON forecast submissions were thus evaluated in two ways that captured different attributes of accuracy and precision: the continuous rank probability score (CRPS) and a CRPS comparison with a baseline (null) model that acted as a benchmark to assess relative gains in forecast skill (Pappenberger et al., 2015; Murphy 1992).

In this study, we analysed a year of submissions to the aquatics theme of the NEON Challenge and assessed how model performance varied among model class, model covariates, and forecast sites. We used the forecast analysis to answer the following research questions: Q1) How does model class and inclusion of covariates affect forecast performance?; Q2) To what extent is relative forecast skill affected by the inclusion of different sources of uncertainty?; and Q3: How consistent are the patterns in forecast performance across sites? We included all Tw and DO forecasts in the analysis of Q1, but focused primarily on Tw forecasts for Q2 and Q3 due to the much higher number of submissions for that variable (see below). To the best of our knowledge, our study is the first analysis that investigates the performance of freshwater

forecasts across multiple model classes, model covariates, and sites using genuine forecasts of the future.

2 Methods

2.1 NEON Challenge Overview

The NEON Challenge has five forecasting themes that cover a range of ecological populations, communities, and ecosystems across the NEON network of monitored freshwater and terrestrial sites. Our co-author team represents a group of the Challenge organisers, cyberinfrastructure developers, and/or forecast submitters.

Submissions were accepted to the aquatics theme of the NEON Challenge starting in 2021 and continuing to the present (>3 years) for forecasts of water quality. Here, we focus on the forecasts of Tw and DO submitted to lake sites within the aquatics theme of the NEON Challenge during 2023, which represented the first full year with sufficient submissions for a robust inter-model comparison.

2.2 Challenge design

2.2.1 NEON data

Water quality data were collected at seven lakes across the US (Figure 1). Tw and DO were collected using in-situ sensors. Full descriptions of the sensors and protocol are included in the data product metadata provided by NEON (DP1.20264.001, NEON TSD) for Tw and DP1.20288.001 for DO (NEON water quality). At each lake, data were only available at one location (generally at the centre, near the deepest point). For the purposes of the Challenge, unpublished data were made available to participants by NEON at a data latency of 2-3 days

after collection. The Tw and DO NEON data products extend back to 2016 but their temporal coverage varies across sites in three ways. First, there is variability in the duration of time-series data available for each site and variable (Figure S1), ranging from 3.1 to 6.6 years (up to 1 January 2023, the beginning of our focal forecasting period). Second, at five lake sites, sensors are removed during winter due to ice formation. Finally, maintenance issues resulted in data gaps at some sites. Consequently, total data availability varied between 167 and 2154 days for each site/variable combination (Figure S1).

2.2.2 Data processing and targets generation

We, as challenge organisers, converted the Tw and DO data supplied by NEON in near-real time to "targets" - observations specific to the challenge - by subsetting the sensor locations, performing additional quality control, and aggregating 30-minute sensor data to daily means. First, the data were subset to include only the surface measurements (top 1 m of the water column). Second, we filtered the data using the existing NEON flags (see metadata) and applied additional quality control measures (e.g., additional filtering for maximum and minimum allowable values for each variable; see Olsson et al. 2024a). The targets data could then be used by teams to calibrate and train models and were used for forecast evaluation.

These processed target data were publicly available to all Challenge teams at a persistent URL location and were updated daily as new data became available. To further support modelling efforts by the teams, we also provided supplementary hourly water temperature profile data collected by NEON at each of the lake sites (derived from NEON DP1.20264.001, see Olsson et al., 2024a). These supplemental data were available to teams to use in model development and training but were not used in forecast evaluation.

2.2.3 Ancillary driver data

NOAA's Global Ensemble Forecasting System (GEFS; Hamill et al., 2022) weather forecast data were made available to forecast teams via functions in the custom R package *neon4cast* (Boettiger & Thomas, 2024). NOAA weather data for all NEON sites were downloaded each day and standardised to be used as driver data and covariates in forecast models. Teams were not required to use weather covariates but providing standardised NOAA weather forecasts ensured that the teams that used weather covariates had consistent data, and weather forecast skill was therefore not the primary driver of differences in aquatic forecast skill among model submissions. Two NOAA data products were used by forecast teams: an ensemble forecast of future weather and a historic weather product. The ensemble weather forecast consisted of 31 ensemble members up to 35 days into the future at each of the 34 sites. The historic product consisted of stacked one day-ahead forecasts from each day as an estimate of observed historical conditions that was consistent with the ensemble weather forecast data available to teams to forecast (i.e., having similar biases, compared to observational weather data) and could be used to calibrate models. Teams were also able to use any other openly-available covariate data in their forecasts, although none chose to do this.

2.2.4 Forecast submission guidelines

Challenge teams were invited to forecast T_w and DO in all of the lakes or in any subset of sites or variables. Forecast submissions were required to have a daily time step of the focal variable(s) over a forecast horizon of at least 1 to 30 days into the future and include an estimate of uncertainty in the forecast. Uncertainty could be represented by submitting a probabilistic

forecast (Gneiting & Katzfuss, 2014), either in the form of a mean and a standard deviation for a normally distributed forecast or as an ensemble forecast for which the uncertainty was represented as a series of predictions that represent a range of future conditions (Gneiting & Katzfuss, 2014). Submissions were required to follow a standardised format (Thomas, et al., 2023; Dietze et al., 2023) to enable automated evaluation and processing. New forecasts were accepted every day and evaluated as new observational data became available (see Section 2.4).

During 2022 and 2023, we ran multiple workshops to introduce the Challenge to a cross-section of aquatic and data scientists and managers to increase forecast submissions to this theme (Meyer et al. 2023; Olsson et al., 2023). In total, more than 300 people attended the workshops in person or online. Workshop materials were also available online for individuals or groups to use independently (Olsson et al., 2023).

2.2.5 Baseline model

Following forecast evaluation best practices (Harris et al., 2018; Lewis et al., 2022), we generated a baseline model that represents a limited (naive) understanding of the system for comparison to the submitted forecast models. It can be helpful to compare submitted forecasts with forecasts generated from baseline models as part of forecast evaluation to identify whether new methods provide additional, useful information beyond uninformed models (Pappenberger et al., 2015; Makridakis et al., 2020; Joliffe & Stephenson, 2012). Specifically, we generated a model that assumes the forecast for a particular day-of-year (DOY) is equal to the mean of historical data on that DOY. The DOY baseline model assumes dynamics will follow the mean conditions for that date in previously observed years (Hyndman & Athanasopoulos, 2021; Jolliffe & Stephenson, 2012). The uncertainty in this DOY forecast was generated by calculating

the standard deviation of the past observations (see Supplementary Information Text S1). The standard deviation of the daily average for the forecast period was used to represent the uncertainty for the whole horizon. The DOY forecast was assumed to follow a normal distribution, given by a mean and standard deviation for each day of year calculated separately for each site and variable.

The baseline model was selected based on the observed dynamics of the variable of interest (Jolliffe & Stephenson, 2012; Pappenberger et al., 2015) as well as being a common baseline for ecological forecasts (e.g., Wheeler et al., 2024; Thomas et al., 2020; Lewis et al., 2022). The DOY model is particularly useful as a baseline when the target variable's dynamics follow a seasonal cycle (Pappenberger et al., 2015), such as variables primarily driven by meteorological forcing. A second baseline model that assumes a forecast is equal to the last observation (persistence; Joliffe & Stephenson, 2012) was also included in submissions but had a lower overall performance for both variables so was not used as a reference.

2.2.6 Forecast evaluation

Forecasts were evaluated against observations using the continuous rank probability score (CRPS), as implemented in the *scoringRules* R package (Jordan et al., 2019). CRPS evaluates the probability distribution of the forecast and assesses both the accuracy and precision of the forecast relative to observations. Specifically, we used a *relative forecast skill* (hereafter, $CRPS_{skill}$) metric to describe how much additional information is gained in each model over a naive model. $CRPS_{skill}$ was calculated based on the difference in CRPS score between the submitted forecast and the DOY baseline model, following Equation 1:

Equation 1. $CRPS_{Skill} = forecast\_score - DOY\_score$

with positive values indicating a submitted forecast showing lower skill relative to the DOY model, and negative values indicating that the DOY model performed better.

2.3 Analyses

We assessed the performance of the forecast models across different horizons and sites by aggregating raw $CRPS_{skill}$ metrics at different temporal and spatial scales. To identify the best performing models per variable, we calculated the mean $CRPS_{skill}$ aggregated across all forecast submission dates, horizons, and sites. To ensure that the comparisons among models were based on a similar number of submissions, we only included models in the analysis that had submissions for 80% of evaluated days (i.e., days with observations). We allowed teams to 'catch-up' their forecasts (i.e., submit forecasts which were not 'real-time' but 'retroactive forecasts' following Joliffe & Stephenson, 2012) when they missed submissions due to any issues with automated cyberinfrastructure. Retroactive forecasts could only use target data and forecasted covariates that would have been available if the forecast was generated in real-time (i.e., a retroactive forecast of water temperature for 1 July 2023 only used observations before this date for model training and was driven by NOAA weather forecasts generated on 30 June 2023 or earlier). No model was represented only by retroactive forecasts. In our analysis, we removed the 16 day-ahead horizon from evaluation because of processing issues when downloading NOAA weather forecasts. The 16 day-ahead horizon had artificially low variance in the forecast that was not present in the other horizons (the 1- to 16 day-ahead forecast becomes available for download from NOAA earlier than the 17- to 35 day-ahead forecast). The processing issue was resolved during the period of evaluation but we excluded the 16 day-ahead horizon regardless so that we could compare forecasts throughout all of 2023.

The reliability of the confidence intervals (CI) was calculated by estimating the number of observations that fell within the confidence intervals specified and thus the degree to which a predicted distribution matches the true underlying distribution of the data. Reliability refers to the statistical agreement of forecast probabilities with observed relative frequencies of events (Gneiting, Balabdaoui & Raftery, 2007; Schepen et al., 2016). A forecast that has perfectly reliable confidence intervals will have the equivalent proportion of the observations falling within the CI (Joliffe & Stephenson, 2012, Thomas et al., 2020): e.g., 80% of observations falling within the 80% confidence interval and 95% of observations falling within the 95% confidence interval. Forecasts with too many observations falling within the CI are said to be 'underconfident', while have too few in the CI is 'overconfident' (as Thomas et al., 2020; Zwart et al., 2020; Ouellet-Proulx et al., 2017).

3 Results

3.1 Forecast inventory

Individuals and teams submitted a total of 100,475 daily forecasts for 1-30 day-ahead horizons using 28 different models to the aquatics lake theme of the NEON Challenge in 2023. Here, we define one forecast as a collection of predictions for 1 to 30 days in the future for a unique combination of forecast starting date, forecast site, forecasted variable, and forecasting model. The 28 models were used in addition to the two baseline models (persistence and DOY models) submitted by Challenge organisers (n = 30 models total). The forecasted variables were unevenly represented in the submissions: 14 models (plus two baselines) were used to submit forecasts for both variables (Tw, DO), 14 models were used to submit forecasts for only Tw, and no models submitted forecasts for only DO (total model submissions for each variable: Tw = 30,

DO = 16). Across all submissions, forecasts of water temperature for the lake sites were most numerous (n = 63,189; 63% of total lake forecasts) and had a greater diversity of model classes and covariates. We note that the 30 Tw models represent only the models that were used to continuously submit forecasts throughout 2023. We omitted 42 other models from one-off forecast submissions that were submitted as part of training tutorials, courses, or workshops; were uploaded for the purposes of model testing; or were submitted by unregistered participants.

The 30 Tw models included a range of model classes and exogenous covariates. The self-reported model classes included empirical models (statistical and time series), machine learning, and process-based models, as well as multi-model ensembles (MME; i.e., predictions were based on an aggregation of other model forecast submissions). Forecast models included a range of exogenous covariates from the NOAA GEFS weather forecasts, with forecasted air temperature being the most commonly used covariate (n = 19, Table S1). No other exogenous covariates (i.e., non-NOAA GEFS weather covariates) were included in any model. Details of all of the models that submitted forecasts in 2023 that met the criteria for inclusion in this analysis are provided in Supplementary Information Text S1.

The 16 DO models represented less diversity in model classes and covariates than the Tw models (Figure 2). The model classes for the DO models included only empirical and ML models (in addition to the baseline models), and air temperature was used as a covariate in six of the 16 DO models (38%).

3.2 How does model class affect forecast performance across all variables?

More Tw forecast models (n = 10) outperformed the baseline than the DO forecast models (n = 6; Figure 2). Only six of the submitted DO models outperformed the DOY baseline

model across all forecast dates and sites (i.e., models had mean positive $CRPS_{skill}$, with a mean relative skill between 0.01 and 0.08 mg/L aggregated across the 1-30 day-ahead horizon (Figure 2c). These six highest performing DO models included both ML and empirical models, of which the highest performing models were ML models that used air temperature as a covariate (Random Forest, Lasso, and XGBoost). The models that did not out-perform the baseline were all empirical, and no PB models were used to forecast DO in lakes.

Unlike DO, the best performing models for water temperature (Tw) were from the full range of model classes (Figure 2a,b). Of the 30 submitted models, ten Tw forecast models outperformed the DOY baseline model when forecasts were aggregated across all sites and horizons for the year of forecasts. Across all sites and forecasts, a PB model had the highest skill (Figure 2a), with a mean $CRPS_{skill}$ of 0.22 °C aggregated across the 1-30 day-ahead horizon. Although the overall top three models were PB models, not all PB models were high performing, as four PB models had a negative mean $CRPS_{skill}$ (Figure 2b).

Altogether, of the different model classes used to submit forecasts of Tw, four of the eight PB models, one of the 13 empirical models, two of the four MME, and all three of the ML models outperformed the baseline DOY model on average over the year (Figure 2a). Machine learning models accounted for three models in the top 10 Tw forecast models, as XGBoost, Random Forest, and Lasso models all had positive $CRPS_{skill}$. Empirical models exhibited the worst performance among the model classes, as only one (the Prophet model, Figure 2a) outperformed the DOY baseline model across all forecasts. Given the higher performance of forecasts for Tw (ten models beating the baseline), as well as the higher diversity of model classes represented in these higher performing models (n = 4), further analyses for addressing Q1, Q2, and Q3 were conducted on the Tw forecasts only.

3.3 Among Tw models, how does model class and inclusion of covariates affect performance across the forecast horizon?

Nine out of the 10 Tw models that outperformed the baseline model included air temperature as a covariate (Figure 2a). The specific inclusion of air temperature as a covariate appeared to confer some skill, as it was not included in any of the five lowest performing models (Figure 2b). However, the inclusion of exogenous covariates did not guarantee high performance of a model, as ten of the models exhibiting negative $CRPS_{skill}$ included air temperature as a covariate, as well as other NOAA weather covariates such as humidity and precipitation (Supplementary Text S1). There was only one model that outperformed the baseline model, the empirical Prophet model, which was based solely on observations and included no exogenous covariates (Figure 2a).

Focusing on Tw, $CRPS_{skill}$ in the most skillful forecasts generally degraded across the forecast horizon (Figure 3a), and, on average, were unable to outperform the DOY baseline at horizons of 15 to 25 days-ahead. The exceptions to this pattern were the Lasso and Random Forest ML models, which showed increases in skill for the first 7-8 days-ahead and then decreases in skill at longer horizons. Generally, the PB models and MME forecasts showed larger rates of degradation compared to the ML and empirical models (Figure 3a). The Prophet ML model exhibited the lowest degradation in skill (0.16 °C to -0.08 °C) across the 30 days, although its skill at 7-16 day-ahead horizons was the lowest of any model that out-performed the baseline (Figure 3a). In comparison, two MME forecasts showed the highest rates of degradation (LER baselines MME and FLARE-LER MME), from high performance at short horizons (0.58 °C and 0.64 °C) to negative relative skill at the longest horizons (-0.24 °C and -0.32 °C). Only one model had positive $CRPS_{skill}$ across the full forecast horizon, the XGBoost ML model, which

had a low rate of skill degradation across the 30 days (0.32 °C, Figure 3a). The models which exhibited negative skill throughout the 30-day forecast horizon generally showed consistently decreasing performance into the future (Figure S2), although the worst performing models had low performance irrespective of forecast horizon.

Out of all Tw models that outperformed the DOY baseline (as determined by the aggregation of skill over the full forecast horizon; Figure 2a), XGBoost had positive relative skill for the full forecast horizon with the FLARE-GLM PB model, and the Random Forest ML model had the next longest durations of positive relative skill (i.e., 19 and 27 days, respectively, over the 30-day forecast horizon), but differed in the timing of these days. The Random Forest model had negative $CRPS_{skill}$ at the start of the forecast horizon and FLARE-GLM had negative relative skill at the end of the forecast period (Figure 3a), although both were only marginally worse-performing than the baseline on the days when their $CRPS_{skill}$ was negative. FLARE-GLM was the most skillful model for the first 16 days of the forecast horizon, dropping only to the 4th highest performer overall at other horizons. In contrast, the best performing model at 30 days ahead, the Random Forest model, was the second worst-performing model at 1-4 days-ahead.

3.4 To what extent is relative forecast skill affected by the inclusion of different sources of uncertainty?

Although submissions were required to include an estimate of forecast uncertainty (Thomas et al., 2023), the sources of uncertainty varied among the models. The most commonly represented source of uncertainty in Tw models was driver uncertainty (n = 22, Table S1), with 13 models including only one source of uncertainty, seven models including two sources, one model including three sources, and eight models including all five sources of uncertainty

(defined in Table 1).

Of the 10 Tw models that had mean positive skill aggregated over the forecast horizon for Tw (Figure 2a), seven included at least three sources of uncertainty and six included five sources (Table 2). All but one model (n = 9) included driver uncertainty (in the form of the NOAA GEFS weather ensembles as covariates), with parameter and process uncertainty the next most common uncertainty source included with these top models (n = 8 models represented this source of uncertainty). In comparison, Tw models that performed less well than the baseline rarely included sources of uncertainty other than driver data uncertainty (Table S1).

The degradation in relative skill for the majority of Tw models at longer horizons was concurrent with an increase in bias (i.e., lower accuracy; Figure 3c) and standard deviation (i.e., lower precision; Figure 3b). The increased relative skill exhibited by two ML models (Lasso and Random Forest) across the first seven days of the forecast horizon (Figure 3a) was concurrent with reductions in absolute bias (Figure 3c). Across the first 10 days, the PB models (FLARE-GLM, FLARE-GOTM) and MMEs that included the PB models (FLARE-LER MME and LER baselines MME) exhibited the lowest absolute bias, which increased steadily across the horizon up to ~20 days ahead. In comparison, the forecast accuracy and to a certain extent, precision, in the Prophet, XGBoost, and Random Forest ML models degraded less, resulting in lower bias and SD at longer horizons (Figure 3c).

Increased standard deviation (i.e., greater uncertainty) across the forecast horizon may indicate a reduction in precision in the forecasts, which can degrade $CRPS_{skill}$ and reliability of the forecast confidence intervals (CI). The top performing Tw models were primarily underconfident (Figure 4a) for the 80% confidence intervals, meaning that >80% of observations fell within the 80% confidence intervals. Generally, the confidence of the forecasts changed little

444 over the horizon, especially beyond the first five days (Figure 4a). Beyond this horizon, only the
445 Random Forest and Lasso ML models showed shifts in confidence beyond 5 days, becoming less
446 overconfident and eventually becoming underconfident at horizons greater than 8 days (Figure
447 4). The XGBoost ML model yielded the most reliable forecasts, with 80.4% of observations in
448 the 80% CI when averaged across horizons (Figure 4). The Prophet model was the only model
449 that out-performed the baseline that was overconfident for the whole forecast horizon, with its
450 uncertainty changing little across the forecast horizon (74-79% of observations in the 80% CI;
451 Figure 4a). The two MME models showed the highest rates of underconfidence, with 91.5% and
452 96.2% points falling on average into the 80% CI (Figure 4). Among the poorer performing Tw
453 models, there was a greater rate of overconfidence, especially at horizons less than 7 days ahead,
454 with 9 out of the 18 models overconfident. The rate of overconfidence increased among all
455 models at the 95% CI (Figure 4b,d), demonstrating poor calibration for models when forecasting
456 observations at the tails of the distribution.
457
458 *3.5 Are the patterns in performance consistent across sites?*
459 Within model classes, Tw forecast $CRPS_{skill}$ showed similar patterns among sites, with
460 the exception of empirical models (Figure 5a). Generally, ML, PB models, and MMEs had
461 positive $CRPS_{skill}$ at PRLA, PRPO, and TOOK, though the latter had a limited number of
462 forecasts given its much shorter buoy deployment duration (Figure S1). In comparison, ML
463 models, PB models, and MMEs generally exhibited negative $CRPS_{skill}$ at SUGG, BARC, and
464 CRAM (Figure 5a).
465 Mean $CRPS_{skill}$ (from the Tw models that outperformed the baseline, as shown in Figure
466 2a) degraded across the forecast horizon for all sites, but remained positive at PRPO and PRLA

for the full 30-day horizon and at TOOK for the first 18 days (Figure 5b). In contrast, at CRAM, LIRO, BARC, and SUGG, $CRPS_{skill}$ was positive between 1 and 12 days-ahead. This high $CRPS_{skill}$ at PRPO, PRLA, and TOOK is likely due to the relative gains against more poorly-performing DOY baseline forecasts at these sites (Figure S3). Focusing on the four months when all lakes had data availability (i.e., when all lakes had buoys deployed) vs. longer time periods did not substantially alter the differences in $CRPS_{skill}$ observed among lakes (Figure S4).

Climate variability may have influenced why some models performed better than others in forecasting out-of-sample conditions. Observations for water temperatures in 2023 show that PRPO and PRLA were warmer than historical conditions represented in the DOY model, especially in May and June (Figure 6). In comparison, CRAM and LIRO, for which models performed worse than the baseline on average, exhibited water temperatures generally within around 2 °C of historical conditions (Figure 6). BARC and SUGG exhibited a smaller range of water temperatures that fell within 2 °C of historical conditions for all months except March (Figure 6).

4 Discussion

Among the 29 models that forecasted water quality variables across seven lakes, 10 models out-performed the baseline model for Tw, and six for DO (Figure 2). Of the 10 best-performing Tw models, there were four PB models that included multiple exogenous weather covariates, three ML models, two multi-model ensembles, and one empirical model, demonstrating that multiple different model classes can yield skillful forecasts for lake water temperature. Our uncertainty analysis showed that poor-performing Tw models were generally more overconfident, likely due to insufficient representation of uncertainty in the forecasts.

Finally, model skill was inconsistent across sites for the best-performing lake temperature forecast models, which may be related to climate variability. Below, we discuss how our findings addressed our research questions, with a focus on the Tw models.

4.1 How do model class and model covariates affect forecast performance?

No individual model submitted to the challenge was the best performing model for both variables, although four models outperformed the baselines for both Tw and DO. These four models – the ML models XGBoost, Random Forest, and Lasso and the empirical model Prophet – show that a range of model types were useful for a range of variable forecasts. High performing models for DO were in both empirical and ML categories, although no PB or MME models were submitted for DO, necessitating further investigation of both model types to potentially improve forecast performance (Olsson et al., 2024c; Hagedorn et al., 2005). In contrast, models outperforming the baseline for Tw came from four model classes (ML, PB, empirical, and MMEs).

In an analysis of Tw models specifically (because of the higher diversity of model classes that were submitted for this variable), we found that PB models that included air temperature as a covariate performed best across all sites (Figure 2a). Air temperature is likely a key covariate for high-performing surface water temperature forecasts because Tw dynamics are primarily driven by processes at the air-water interface of lakes (Schmid & Read, 2021; Piccolroaz et al., 2024), with air temperature a causal forcing variable (Livingstone & Padisák, 1989). PB models that used additional meteorological parameters (e.g., incoming short-wave radiation, relative humidity, wind speed) to calculate heat fluxes to mechanistically derive water temperatures resulted in even higher performing forecasts (Figure 2). One exception was a simple-physics PB

model that had insufficient uncertainty representation and thus was not able to out-perform the baseline model (Supplementary Text S1). Altogether, our results strongly support that including the dominant drivers of water temperature (namely, air temperature) not surprisingly improved the performance of lake water temperature forecasts.

In contrast to the Tw PB models, the domain-agnostic models (i.e., models that do not include any mechanistic information about lake functioning; ML and empirical models) showed less degradation across the forecast horizon, which may be potentially due to the non-dynamic nature of the methods (Supplementary Table S1). In comparison, the PB models were more skillful at short horizons, suggesting that forecasters might choose different Tw models based on the horizon needed. XGBoost, Lasso, and Random Forest ML models and the empirical Prophet model were less skillful than the PB models and PB-MMEs in the first 10 days, but become more skillful than the PB models at longer horizons due to their low rates of degradation. XGBoost was the only model to have a positive skill across the full forecast horizon (on average for all forecasts and sites), highlighting a robust method for forecasting Tw at any site in our study. Our results are similar to other ecological forecasting studies: for example, domain-agnostic models outperformed PB models in a penguin population forecasting competition in which annual populations were forecasted up to 3 years ahead (Humphries et al., 2019). Similarly, simple time-series models have shown promise in other ecological population forecasts (Ward et al., 2014). In the NEON Challenge, the same ML and empirical models that performed well for Tw also performed well for DO forecasts, on average out-performing the DOY baseline, and thereby representing robust methods across multiple variables.

Reduction in skill of Tw forecasts over the forecast horizon may be linked to a reduction in skill of the air temperature forecasts being used as model driver data. The Prophet model,

which was the only model that out-performed the baseline that did not include air temperature as a covariate (or any covariates at all), showed less degradation in forecast performance than the overall better performing PB models, although this represents only a single model. The PB models, generally, benefit from high weather forecast skill at shorter horizons (Petchey et al., 2015, Zhou et al., 2021), but degrade in performance along with the performance of their covariates. Beyond 10 days-ahead, when the weather forecasts are less skillful (Zhou et al., 2021), the PB models performance suffered, suggesting that forecasters seeking to optimise performance at longer horizons should focus on models that are less dependent upon meteorological driver data (e.g., time series models).

The differences in the forecast horizons at which each Tw model was most skillful may present opportunities for generating MMEs or hybrid models (e.g., combining domain-agnostic models with PB models) to exploit the strengths of multiple model types across the forecast horizon. Hybrid model approaches have shown high performance in other forecasting challenges and competitions (Makridakis et al., 2020; Clark et al., 2022) and MMEs are most successful when the individual model structures are more diverse (Olsson et al., 2024c; Petropoulos et al., 2022; Dormann et al., 2018). The performance of the MMEs in this NEON Challenge synthesis was not consistent with previous studies and other forecasting challenges, in which MMEs showed the best performance (Makridakis et al., 2020; Clark et al., 2022). For example, in forecasts of tick disease incidence, the simple model average of four individual models was better than any individual model (Clark et al., 2022), and the winner of the M4 forecasting competition (a wide-ranging timeseries forecasting challenge) as a combination of statistical and empirical models (Makridakis et al., 2020). Similarly, in a recent single-site lake study, forecasts generated by an MME composed of three PB and two baseline models outperformed the

559 individual models across two years (Olsson et al., 2024c). Conversely, in this analysis, the same
560 MME had lower relative skill, higher bias, and higher uncertainty than some of the individual
561 models from which it was derived (Figure 2). This discrepancy in MME performance could be
562 caused by poor calibration in the individual models at some of the lake sites. The individual
563 models included in this study were almost all underconfident (Figure 4), which resulted in very
564 large uncertainty in the MMEs and likely contributed to their poor performance, as MME
565 forecasts have been shown to be most successful when the individual constituent models are
566 slightly overconfident (Wang et al., 2022; Hagedorn et al., 2005). Methods such as trimming,
567 where distributions are narrowed, could help constrain MME uncertainty, increasing the overall
568 skill of these forecasts (Howerton et al., 2023).
569
570 *4.2 To what extent is relative forecast skill affected by the inclusion of different sources of*
571 *uncertainty?*
572 Our synthesis suggests that representation of forecast uncertainty is important for
573 determining the overall forecast performance of probabilistic Tw forecasts. The top performing
574 Tw models often included multiple sources of uncertainty (up to $n = 5$, Table 2), unlike the lower
575 performing models, which frequently only included driver uncertainty. Consequently, many
576 poor-performing models were overconfident in their predictions, suggesting there was
577 insufficient uncertainty included in those forecasts, especially at shorter horizons (Figure 4).
578 These results suggest that driver uncertainty alone is not a sufficient representation of the total
579 uncertainty, especially given that weather forecasts are themselves often overconfident at these
580 horizons (Zhou et al., 2022). When these weather forecasts are used as driver data for overfitted
581 lake models (Zwart et al., 2023), overconfidence in water quality forecasts is even more likely to

582 occur. Overconfidence of forecasts was also reported in a forest phenology forecast synthesis, in
583 which forecasts that included covariates were overconfident at shorter horizons (Wheeler et al.,
584 2024). In our analysis, the Lasso and Random Forest ML models, which only included driver
585 uncertainty, showed performance improvements at longer horizons as the uncertainty from the
586 weather forecasts increased and the water temperature forecasts became less overconfident
587 (Figure 4). Furthermore, the ML XGBoost model, which included process uncertainty in addition
588 to driver uncertainty, outperformed the other ML models at shorter horizons.

589 Improving the representation of uncertainty in forecasts, as quantified by the reliability of
590 forecast confidence intervals, is important for management (Ramos et al., 2013; Crochemore et
591 al., 2021). Use of ecological forecasts by decision makers is likely to improve if forecast
592 uncertainty is well quantified and confidence intervals are appropriate (Ramos et al., 2013;
593 Buizza 2008; Nadav-Greensberg & Joslyn, 2009). Underconfidence and overconfidence limit the
594 use of forecasts for management, as underconfident forecasts provide too wide of a range of
595 potential future conditions and overconfident forecasts underestimate the possible range of
596 conditions, with both leading to inappropriate management actions (Crochemore et al., 2021).
597 Consequently, our results suggest that including more than one source of uncertainty may help
598 increase the usability of forecasts as decision support tools.

599

600 *4.3 Is model forecast performance consistent across sites?*

601 Tw forecast performance varied among sites, with the relative gain in skill likely due to
602 the lower performance of baseline models at some lakes, especially at PRPO and PRLA, two
603 lakes in North Dakota. The DOY baseline model had the lowest performance at PRPO and
604 PRLA, potentially because 2023 conditions in these two lakes were substantially different from

historical observations, resulting in a lower performing baseline forecast (Figures 6, S3). This is consistent with a previous single-model forecasting study (FLARE-GLM) that also showed improved performance above a DOY baseline for these two sites, especially at shorter horizons (Thomas et al., 2023). Differences from historical conditions that exceeded 3 °C resulted in poor DOY baseline performance in that study. Our results suggest that if there is a divergence of water temperature of this magnitude, using a PB or ML model provides a much stronger forecasting approach than a baseline model. All model classes except the empirical model class showed positive skill compared to the DOY baseline at PRLA and PRPO as well as at TOOK, to a lesser extent. As environmental conditions further exceed historical means due to global change, models that only consider patterns from long-term historical observations may be less valuable than models that are able to infer ecological processes or use recently-observed data in generating forecasts.

4.4 Value and refinements for forecasting challenges

Forecasting challenges provide a compelling opportunity to learn about ecological predictability over gradients of time, space, ecological level of organization, and forecasting methods. The submissions from 30 models (including two baselines) to the aquatics lake theme of the NEON Challenge covered a range of model classes and approaches. However, since the NEON Challenge was open to the community and we did not specifically guide the types of submissions, the breadth of models was not exhaustive and therefore some questions remain. Specifically, quantifying the value of different covariates to different models (e.g., XGBoost, linear models, Random Forest) would be best done by comparing forecasts with the same modelling approach but with differing covariates and quantitatively seeing how forecast skill

changes with their addition or removal. It is possible that this 'model selection' was done by teams before forecasts were submitted and that the final model submitted to the Challenge was the optimal structure, but we cannot know from the submitted metadata whether these models represent each team's "best" attempt at producing a forecast.

We also saw uneven representation in the variables being forecasted, with more submitted forecasts of Tw than DO. We identified several potential factors that contributed to this uneven representation. First, NEON Challenge training materials were focused on lake temperature forecasting, which may have skewed submissions to this variable because participants in workshops may have been more likely to modify pre-existing code for submitting a new model type to Tw, rather than develop new code for DO submissions. Second, water temperature may have been an easier, more "introductory" forecast target variable as there are well-established mechanistic processes linked to driver datasets (e.g., meteorology) that were made readily available for teams to use. Conversely, the drivers of DO concentrations are much more complex, drawing from physical, chemical, and biological processes (Langman et al., 2010, Carey, 2023; Hanson et al., 2006). Additional driver data needed to model lake DO processes, such as nutrients and inflows, were not as easily accessible as the historical and forecasted meteorological drivers from the Challenge organisers. Overall, the best performing models for lake water temperature are unlikely to be optimal for a wide range of other variables and ecosystems, motivating future work and the need for more submissions to the NEON Challenge to understand how their forecast model performance varies across lake variables.

The NEON Challenge aims to both provide training for the larger ecological community and enable quantifying the fundamental predictability of ecological variables. While the submitted forecasts have limitations relative to a standardised modelling exercise (e.g., a

formalised multi-model intercomparison project), the Challenge trained >300 individuals and teams, resulted in the development of novel ecological models, and introduced forecasting to a broad community of researchers. The NEON Challenge recruited participants as a fun, training-focused, educational, and accessible opportunity to learn forecasting in a low-stakes environment, with flexible deadlines and registration. Therefore, despite the potential drawbacks of an open forecasting challenge (vs. a standardized multi-model comparison), it is worth noting that even the forecasts that were omitted from the final analysis still represent participants who may have never generated even a single ecological forecast without the inclusive atmosphere of the NEON Challenge and its training materials.

The NEON Challenge also sets the stage for future forecasting model analyses. For example, future work could address whether the inclusion of exogenous covariates in models produces forecasts that are overconfident at shorter horizons for other ecological variables, which could be corrected using multiple sources of uncertainty. Similarly, it would be useful to investigate whether the domain-agnostic models that outperformed the baseline for DO and Tw perform similarly well when forecasting other ecological variables. The spatial and temporal extent of NEON data, as well as the range of ecological variables on which data are collected, provides a suite of opportunities to continue to investigate these questions and as a platform to grow the field of ecological forecasting.

5 Conclusion

Our synthesis of >100,000 submissions to the NEON Forecasting Challenge demonstrates that a number of model classes were able to out-perform a DOY baseline model to forecast water temperature and dissolved oxygen across seven lake sites, providing insight into

optimal forecasting approaches for different contexts. Water temperature models that included air temperature as an exogenous covariate and those that included multiple sources of uncertainty generally performed well and came from process-based, empirical, machine learning, and multi-model ensemble model classes. The relative skill of these models was shown to be highest at sites that exhibited conditions outside of historical observations. These forecasting methods are likely to become increasingly valuable for guiding decision-making in a world in which ecosystems are become more variable and continue to move outside of historically observed conditions. Overall, our results highlight the value of forecasting challenges to advance the development of ecological forecasts for both theory and management.

Acknowledgments

The authors wish to thank the Ecological Forecasting Initiative Research Coordination Network, funded by the U.S. National Science Foundation (NSF) grant DEB-1926388, and the original aquatics theme design team members for their work in providing a foundation for the ongoing NEON Challenge. We also wish to thank NEON staff, especially Bobby Hensley and Kaelin Cawley, for their help and guidance obtaining and troubleshooting aquatics data for the Challenge. FO was financially supported by NSF grant DBI-1933016; CCC and MEL were supported by DEB-1926050; RQT was financially supported by OAC-2209866; and ASL was financially supported by an NSF graduate research fellowship (DGE-1840995), DEB-1753639, and funding from the Virginia Tech College of Science Roundtable.

695 **Authorship contribution statement**

696 RQT, FO, and CCC designed and developed the NEON EFI Aquatics Challenge. RQT, CB, and
697 FO developed cyberinfrastructure. All co-authors contributed forecasts. FO, CCC, and RQT
698 developed the synthesis and analysis approach with feedback from all co-authors. FO led
699 manuscript writing, supported by CCC and RQT. All co-authors contributed text describing
700 forecasting approaches, provided feedback, and approved the final manuscript.

701

702 **Conflict of interest statement**

703 The authors report no conflicts of interest.

References

Almeida, M. C., Y. Shevchuk, G. Kirillin, P. M. M. Soares, R. M. Cardoso, J. P. Matos, R. M. Rebelo, A. C. Rodrigues, and P. S. Coelho. 2022. Modeling reservoir surface temperatures for regional and global climate models: a multi-model study on the inflow and level variation effects. Geoscientific Model Development:137–197.

Biggerstaff, M., M. Johansson, D. Alper, L. C. Brooks, P. Chakraborty, D. C. Farrow, S. Hyun, S. Kandula, C. McGowan, N. Ramakrishnan, R. Rosenfeld, J. Shaman, R. Tibshirani, R. J. Tibshirani, A. Vespignani, W. Yang, Q. Zhang, and C. Reed. 2018. Results from the second year of a collaborative effort to forecast influenza seasons in the United States. Epidemics 24:26–33.

Boettiger, C., and R. Q. Thomas. 2024. neon4cast: Helper utilities for the EFI NEON Forecast Challenge. R package version 0.1.0.

Bojer, C. S., and J. P. Meldgaard. 2021. Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting 37:587–603.

Bradford, J. B., J. L. Betancourt, B. J. Butterfield, S. M. Munson, and T. E. Wood. 2018. Anticipatory natural resource science and management for a changing future. Frontiers in Ecology and the Environment 16:295–303.

Buizza, R. 2008. The value of probabilistic prediction. Atmospheric Science Letters 9:36–42.

Caissie, D., M. E. Thistle, and L. Benyahya. 2017. River temperature forecasting: case study for Little Southwest Miramichi River (New Brunswick, Canada). Hydrological Sciences Journal 62:683–697.

Carey, C. C. 2023. Causes and consequences of changing oxygen availability in lakes: Kilham Plenary Lecture Article. Inland Waters 13:316–326.

727 Carey, C. C., W. M. Woelmer, M. E. Lofton, R. J. Figueiredo, B. J. Bookout, R. S. Corrigan, V.
728 Daneshmand, A. G. Hounshell, D. W. Howard, A. S. L. Lewis, R. P. McClure, H. L.
729 Wander, N. K. Ward, and R. Q. Thomas. 2022. Advancing lake and reservoir water
730 quality management with near-term, iterative ecological forecasting. Inland Waters
731 12:107–120.
732 Carrizo, S. F., S. C. Jähnig, V. Bremerich, J. Freyhof, I. Harrison, F. He, S. D. Langhans, K.
733 Tockner, C. Zarfl, and W. Darwall. 2017. Freshwater Megafauna: Flagships for
734 Freshwater Biodiversity under Threat. BioScience 67:919–927.
735 Cheng, M., F. Fang, T. Kinouchi, I. M. Navon, and C. C. Pain. 2020. Long lead-time daily and
736 monthly streamflow forecasting using machine learning methods. Journal of Hydrology
737 590:125376.
738 Clark, N. J., T. Proboste, G. Weerasinghe, and R. J. S. Magalhães. 2022. Near-term forecasting
739 of companion animal tick paralysis incidence: An iterative ensemble model. PLoS
740 Computational Biology 18:e1009874.
741 Clayer, F., L. Jackson-Blake, D. Mercado-bettín, M. Shikhani, A. French, D. Mercado, M.
742 Shikhani, A. French, T. Moore, J. Sample, M. Norling, M.-D. Frias, S. Herrera, E. De
743 Eyto, E. Jennings, K. Rinke, L. Van Der Linden, and R. Marcé. 2023. Sources of skill in
744 lake temperature, discharge and ice-off seasonal forecasting tools. Hydrology and Earth
745 System Sciences 27:1361–1381.
746 Crochemore, L., C. Cantone, I. G. Pechlivanidis, and C. S. Photiadou. 2021. How Does Seasonal
747 Forecast Performance Influence Decision-Making? Insights from a Serious Game.
748 Bulletin of the American Meteorological Society 102:E1682–E1699.

749 Di Nunno, F., S. Zhu, M. Ptak, M. Sojka, and F. Granata. 2023. A stacked machine learning
750 model for multi-step ahead prediction of lake surface water temperature. Science of The
751 Total Environment 890:164323.
752 Dietze, M. C. 2017. Prediction in ecology: a first-principles framework. Ecological Applications
753 27:2048–2060.
754 Dietze, M. C., A. Fox, L. M. Beck-Johnson, J. L. Betancourt, M. B. Hooten, C. S. Jarnevich, T.
755 H. Keitt, M. A. Kenney, C. M. Laney, L. G. Larsen, H. W. Loescher, C. K. Lunch, B. C.
756 Pijanowski, J. T. Randerson, E. K. Read, A. T. Tredennick, R. Vargas, K. C. Weathers,
757 and E. P. White. 2018. Iterative near-term ecological forecasting: Needs, opportunities,
758 and challenges. Proceedings of the National Academy of Sciences 115:1424–1432.
759 Dormann, C. F., J. M. Calabrese, G. Guillera-Arroita, E. Matechou, V. Bahn, K. Bartoń, C. M.
760 Beale, S. Ciuti, J. Elith, K. Gerstner, J. Guelat, P. Keil, J. J. Lahoz-Monfort, L. J. Pollock,
761 B. Reineking, D. R. Roberts, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, S. N.
762 Wood, R. O. Wüest, and F. Hartig. 2018. Model averaging in ecology: a review of
763 Bayesian, information-theoretic, and tactical approaches for predictive inference.
764 Ecological Monographs 88:485–504.
765 Dudgeon, D., A. H. Arthington, M. O. Gessner, Z. I. Kawabata, D. J. Knowler, C. Lévêque, R. J.
766 Naiman, A. H. Prieur-Richard, D. Soto, M. L. J. Stiassny, and C. A. Sullivan. 2006.
767 Freshwater biodiversity: Importance, threats, status and conservation challenges.
768 Biological Reviews of the Cambridge Philosophical Society 81:163–182.
769 Fer, I., A. K. Gardella, A. N. Shiklomanov, E. E. Campbell, E. M. Cowdery, M. G. De Kauwe,
770 A. Desai, M. J. Duveneck, J. B. Fisher, K. D. Haynes, F. M. Hoffman, M. R. Johnston, R.
771 Kooper, D. S. LeBauer, J. Mantooth, W. J. Parton, B. Poulter, T. Quaife, A. Raiho, K.

772 Schaefer, S. P. Serbin, J. Simkins, K. R. Wilcox, T. Viskari, and M. C. Dietze. 2021.
773 Beyond ecosystem modeling: A roadmap to community cyberinfrastructure for
774 ecological data-model integration. Global Change Biology 27:13–26.
775 Gneiting, T., F. Balabdaoui, and A. E. Raftery. 2007. Probabilistic Forecasts, Calibration and
776 Sharpness. Journal of the Royal Statistical Society Series B: Statistical Methodology
777 69:243–268.
778 Gneiting, T., and M. Katzfuss. 2014. Probabilistic Forecasting. Annual Review of Statistics and
779 Its Application 1:125–151.
780 Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer. 2005. The rationale behind the success of
781 multi-model ensembles in seasonal forecasting – I. Basic concept. Tellus A: Dynamic
782 Meteorology and Oceanography 57:219.
783 Hamill, T. M., J. S. Whitaker, A. Shlyaeva, G. Bates, S. Fredrick, P. Pegion, E. Sinsky, Y. Zhu,
784 V. Tallapragada, H. Guan, X. Zhou, and J. Woollen. 2022. The Reanalysis for the Global
785 Ensemble Forecast System, Version 12. Monthly Weather Review 150:59–79.
786 Hanson, P. C., S. R. Carpenter, D. E. Armstrong, E. H. Stanley, and T. K. Kratz. 2006. Lake
787 dissolved inorganic carbon and dissolved oxygen: Changing drivers from days to
788 decades. Ecological Monographs 76:343–363.
789 Harris, D. J., S. D. Taylor, and E. P. White. 2018. Forecasting biodiversity in breeding birds
790 using best practices. PeerJ 6:e4278.
791 Hipsey, M. R., L. C. Bruce, C. Boon, B. Busch, C. C. Carey, D. P. Hamilton, P. C. Hanson, J. S.
792 Read, E. De Sousa, M. Weber, and L. A. Winslow. 2019. A General Lake Model (GLM
793 3.0) for linking with high-frequency sensor data from the Global Lake Ecological
794 Observatory Network (GLEON). Geoscientific Model Development 12:473–523.

795 Howerton, E., M. C. Runge, T. L. Bogich, R. K. Borchering, H. Inamine, J. Lessler, L. C.
796 Mullany, W. J. M. Probert, C. P. Smith, S. Truelove, C. Viboud, and K. Shea. 2023.
797 Context-dependent representation of within- and between-model uncertainty: aggregating
798 probabilistic predictions in infectious disease epidemiology. Journal of The Royal
799 Society Interface 20:20220659.
800 Huang, B., C. Langpap, and R. M. Adams. 2011. Using instream water temperature forecasts for
801 fisheries management: An application in the pacific northwest. Journal of the American
802 Water Resources Association 47:861–876.
803 Humphries, G. R. W., C. Che-Castaldo, P. J. Bull, G. Lipstein, A. Ravia, B. Carrión, T. Bolton,
804 A. Ganguly, and H. J. Lynch. 2018. Predicting the future is hard and other lessons from a
805 population time series data science competition. Ecological Informatics 48:1–11.
806 Hyndman, R. J., and G. Athanasopoulos. 2021. Forecasting: principles and practice. OTexts,
807 Melbourne, Australia. OTexts.com/fpp3.
808 IPCC (Intergovernmental Panel on Climate Change). 2023. Technical Summary. Page Climate
809 Change 2021 – The Physical Science Basis. Cambridge University Press.
810 Johansson, M. A., K. M. Apfeldorf, S. Dobson, J. Devita, A. L. Buczak, B. Baugher, L. J. Moniz,
811 T. Bagley, S. M. Babin, E. Guven, T. K. Yamana, J. Shaman, T. Moschou, N. Lothian, A.
812 Lane, G. Osborne, G. Jiang, L. C. Brooks, D. C. Farrow, S. Hyun, R. J. Tibshirani, R.
813 Rosenfeld, J. Lessler, N. G. Reich, D. A. T. Cummings, S. A. Lauer, S. M. Moore, H. E.
814 Clapham, R. Lowe, T. C. Bailey, M. García-Díez, M. S. Carvalho, X. Rodó, T. Sardar, R.
815 Paul, E. L. Ray, K. Sakrejda, A. C. Brown, X. Meng, O. Osoba, R. Vardavas, D.
816 Manheim, M. Moore, D. M. Rao, T. C. Porco, S. Ackley, F. Liu, L. Worden, M.
817 Convertino, Y. Liu, A. Reddy, E. Ortiz, J. Rivero, H. Brito, A. Juarrero, L. R. Johnson, R.

B. Gramacy, J. M. Cohen, E. A. Mordecai, C. C. Murdock, J. R. Rohr, S. J. Ryan, A. M. Stewart-Ibarra, D. P. Weikel, A. Jutla, R. Khan, M. Poultney, R. R. Colwell, B. Rivera-García, C. M. Barker, J. E. Bell, M. Biggerstaff, D. Swerdlow, L. Mier-Y-Teran-Romero, B. M. Forshey, J. Trtanj, J. Asher, M. Clay, H. S. Margolis, A. M. Hebbeler, D. George, and J. P. Chretien. 2019. An open challenge to advance probabilistic forecasting for dengue epidemics. Proceedings of the National Academy of Sciences of the United States of America 116:24268–24274.

Jolliffe, I. T., and D. B. Stephenson. 2012. Forecast verification: a practitioner's guide in atmospheric science. Page (I. T. Jolliffe and D. B. Stephenson, Eds.). 2nd edition. Wiley Blackwell, Oxford.

Langman, O., P. Hanson, S. Carpenter, and Y. Hu. 2010. Control of dissolved oxygen in northern temperate lakes over scales ranging from minutes to days. Aquatic Biology 9:193–202.

Lewis, A. S. L., C. R. Rollinson, A. J. Allyn, J. Ashander, S. Brodie, C. B. Brookson, E. Collins, M. C. Dietze, A. S. Gallinat, N. Juvigny-Khenafou, G. Koren, D. J. McGlinn, H. Moustahfid, J. A. Peters, N. R. Record, C. J. Robbins, J. Tonkin, and G. M. Wardle. 2023. The power of forecasts to advance ecological theory. Methods in Ecology and Evolution 14:746–756.

Lewis, A. S. L. L., W. M. Woelmer, H. L. Wander, D. W. Howard, J. W. Smith, R. P. McClure, M. E. Lofton, N. W. Hammond, R. S. Corrigan, R. Q. Thomas, and C. C. Carey. 2022. Increased adoption of best practices in ecological forecasting enables comparisons of forecastability. Ecological Applications 32:e02500.

Liu, X., J. Feng, and Y. Wang. 2019. Chlorophyll a predictability and relative importance of factors governing lake phytoplankton at different timescales. Science of the Total Environment 648:472–480.

Livingstone, D. M., and J. Padisák. 2007. Large-scale coherence in the response of lake surface-water temperatures to synoptic-scale climate forcing during summer. Limnology and Oceanography 52:896–902.

Loescher, H. W., E. F. Kelly, and R. Lea. 2017. National Ecological Observatory Network: Beginnings, Programmatic and Scientific Challenges, and Ecological Forecasting. Terrestrial Ecosystem Research Infrastructures: Challenges and Opportunities:27–52.

Makridakis, S., E. Spiliotis, and V. Assimakopoulos. 2020. The M4 Competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting 36:54–74.

Michener, W. K., and M. B. Jones. 2012. Ecoinformatics: Supporting ecology as a data-intensive science. Trends in Ecology and Evolution 27:85–93.

Murphy, A. H. 1992. Climatology, Persistence, and Their Linear Combination as Standards of Reference in Skill Scores. Weather and Forecasting 7:692–698.

Mylne, K. R. 2002. Decision-making from probability forecasts based on forecast value. Meteorological Applications 9:307–315.

Nadav-Greenberg, L., and S. L. Joslyn. 2009. Uncertainty Forecasts Improve Decision Making Among Nonexperts. Journal of Cognitive Engineering and Decision Making 3:209–227.

Olsson, F., C. C. Carey, C. Boettiger, G. Harrison, R. Ladwig, M. Lapeyrolerie, A. S. L. Lewis, M. E. Lofton, F. Montetealegre-Mora, J. S. Rabaey, C. J. Robbins, X. Yang, and R. Q. Thomas. 2024a. What can we learn from 100,000 freshwater forecasts? A synthesis from

the NEON Ecological Forecasting Challenge: scores and targets. Zenodo. https://doi.org/10.5281/zenodo.11087208

Olsson, F., T. N. Moore, C. C. Carey, A. Breef-Pilz, and R. Q. Thomas. 2024c. A Multi-Model Ensemble of Baseline and Process-Based Models Improves the Predictive Skill of Near-Term Lake Forecasts. Water Resources Research 60.

Olsson, F., R. Q. Thomas, C. Boettiger, and M. E. Lofton. 2023. OlssonF/NEON-forecast-challenge-workshop: EFI NEON Forecast Challenge Workshop (v.1.1.0). Zenodo. https://doi.org/10.5281/zenodo.8316966

Olsson, F., R. Q. Thomas, and C. C. Carey. 2024b. What can we learn from 100,000 freshwater forecasts? A synthesis from the NEON Ecological Forecasting Challenge: scripts. Zenodo. https://doi.org/10.5281/zenodo.11093206

Ouellet-Proulx, S., A. St-Hilaire, and M. A. Boucher. 2017. Water temperature ensemble forecasts: Implementation using the CEQUEAU model on two contrasted river systems. Water (Switzerland) 9.

Page, T., P. J. Smith, K. J. Beven, I. D. Jones, J. A. Elliott, S. C. Maberly, E. B. Mackay, M. De Ville, and H. Feuchtmayr. 2018. Adaptive forecasting of phytoplankton communities. Water Research 134:74–85.

Pappenberger, F., M. H. Ramos, H. L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller, and P. Salamon. 2015. How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. Journal of Hydrology 522:697–713.

Petchey, O. L., M. Pontarp, T. M. Massie, S. Kéfi, A. Ozgul, M. Weilenmann, G. M. Palamara, F. Altermatt, B. Matthews, J. M. Levine, D. Z. Childs, B. J. Mcgill, M. E. Schaepman, B. Schmid, P. Spaak, A. P. Beckerman, F. Pennekamp, and I. S. Pearse. 2015. The

ecological forecast horizon, and examples of its uses and determinants. Ecology Letters 18:597–611.

Petropoulos, F., D. Apiletti, V. Assimakopoulos, M. Z. Babai, D. K. Barrow, S. Ben Taieb, C. Bergmeir, R. J. Bessa, J. Bijak, J. E. Boylan, J. Browell, C. Carnevale, J. L. Castle, P. Cirillo, M. P. Clements, C. Cordeiro, F. L. Cyrino Oliveira, S. De Baets, A. Dokumentov, J. Ellison, P. Fiszeder, P. H. Franses, D. T. Frazier, M. Gilliland, M. S. Gönül, P. Goodwin, L. Grossi, Y. Grushka-Cockayne, M. Guidolin, M. Guidolin, U. Gunter, X. Guo, R. Guseo, N. Harvey, D. F. Hendry, R. Hollyman, T. Januschowski, J. Jeon, V. R. R. Jose, Y. Kang, A. B. Koehler, S. Kolassa, N. Kourentzes, S. Leva, F. Li, K. Litsiou, S. Makridakis, G. M. Martin, A. B. Martinez, S. Meeran, T. Modis, K. Nikolopoulos, D. Önkal, A. Paccagnini, A. Panagiotelis, I. Panapakidis, J. M. Pavía, M. Pedio, D. J. Pedregal, P. Pinson, P. Ramos, D. E. Rapach, J. J. Reade, B. Rostami-Tabar, M. Rubaszek, G. Sermpinis, H. L. Shang, E. Spiliotis, A. A. Syntetos, P. D. Talagala, T. S. Talagala, L. Tashman, D. Thomakos, T. Thorarinsdottir, E. Todini, J. R. Trapero Arenas, X. Wang, R. L. Winkler, A. Yusupova, and F. Ziel. 2022. Forecasting: theory and practice. International Journal of Forecasting 38:705–871.

Piccolroaz, S., M. Toffolon, and B. Majone. 2013. A simple lumped model to convert air temperature into surface water temperature in lakes. Hydrology and Earth System Sciences 17:3323–3338.

Piccolroaz, S., S. Zhu, R. Ladwig, L. Carrea, S. Oliver, A. P. Piotrowski, M. Ptak, R. Shinohara, M. Sojka, R. I. Woolway, and D. Z. Zhu. 2024. Lake Water Temperature Modeling in an Era of Climate Change: Data Sources, Models, and Future Prospects. Reviews of Geophysics 62.

908 Qu, B., X. Zhang, F. Pappenberger, T. Zhang, and Y. Fang. 2017. Multi-Model Grand Ensemble
909 Hydrologic Forecasting in the Fu River Basin Using Bayesian Model Averaging. Water
910 9:74.
911 Ramos, M. H., S. J. Van Andel, and F. Pappenberger. 2013. Do probabilistic forecasts lead to
912 better decisions? Hydrology and Earth System Sciences 17:2219–2232.
913 Read, J. S., X. Jia, J. Willard, A. P. Appling, J. A. Zwart, S. K. Oliver, A. Karpatne, G. J. A.
914 Hansen, P. C. Hanson, W. Watkins, M. Steinbach, and V. Kumar. 2019. Process-Guided
915 Deep Learning predictions of lake water temperature. Water Resources Research
916 55:9173–9190.
917 Reid, A. J., A. K. Carlson, I. F. Creed, E. J. Eliason, P. A. Gell, P. T. J. Johnson, K. A. Kidd, T.
918 J. MacCormack, J. D. Olden, S. J. Ormerod, J. P. Smol, W. W. Taylor, K. Tockner, J. C.
919 Vermaire, D. Dudgeon, and S. J. Cooke. 2019. Emerging threats and persistent
920 conservation challenges for freshwater biodiversity. Biological Reviews 94:849–873.
921 Richardson, D. C., A. Filazzola, R. I. Woolway, M. A. Imrit, D. Bouffard, G. A. Weyhenmeyer,
922 J. Magnuson, and S. Sharma. 2024. Nonlinear responses in interannual variability of lake
923 ice to climate change. Limnology and Oceanography 69:789–801.
924 Rousso, B. Z., E. Bertone, R. Stewart, and D. P. Hamilton. 2020. A systematic literature review
925 of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. Water
926 Research 182:115959.
927 Saber, A., D. E. James, and D. F. Hayes. 2020. Long-term forecast of water temperature and
928 dissolved oxygen profiles in deep lakes using artificial neural networks conjugated with
929 wavelet transform. Limnology and Oceanography 65:1297–1317.

930 Schepen, A., T. Zhao, Q. J. Wang, S. Zhou, and P. Feikema. 2016. Optimising seasonal
931 streamflow forecast lead time for operational decision making in Australia. Hydrology
932 and Earth System Sciences 20:4117–4128.
933 Siam, M. S., and E. A. B. Eltahir. 2017. Climate change enhances interannual variability of the
934 Nile river flow. Nature Climate Change 7:350–354.
935 Thomas, R. Q., C. Boettiger, C. C. Carey, M. C. Dietze, L. R. Johnson, M. A. Kenney, J. S.
936 McLachlan, J. A. Peters, E. R. Sokol, J. F. Weltzin, A. Willson, and W. M. Woelmer.
937 2023. The NEON Ecological Forecasting Challenge. Frontiers in Ecology and the
938 Environment 21:112–113.
939 Tulloch, A. I. T., V. Hagger, and A. C. Greenville. 2020. Ecological forecasts to inform near-
940 term management of threats to biodiversity. Global Change Biology 26:5816–5828.
941 Viboud, C., K. Sun, R. Gaffey, M. Ajelli, L. Fumanelli, S. Merler, Q. Zhang, G. Chowell, L.
942 Simonsen, and A. Vespignani. 2018. The RAPIDD ebola forecasting challenge: Synthesis
943 and lessons learnt. Epidemics 22:13–21.
944 Wang, X., R. J. Hyndman, F. Li, and Y. Kang. 2023. Forecast combinations: An over 50-year
945 review. International Journal of Forecasting 39:1518–1547.
946 Ward, E. J., E. E. Holmes, J. T. Thorson, and B. Collen. 2014. Complexity is costly: a meta-
947 analysis of parametric and non-parametric methods for short-term population forecasting.
948 Oikos 123:652–661.
949 Wheeler, K. I., M. C. Dietze, D. LeBauer, J. A. Peters, A. D. Richardson, A. A. Ross, R. Q.
950 Thomas, K. Zhu, U. Bhat, S. Munch, R. F. Buzbee, M. Chen, B. Goldstein, J. Guo, D.
951 Hao, C. Jones, M. Kelly-Fair, H. Liu, C. Malmborg, N. Neupane, D. Pal, V. Shirey, Y.
952 Song, M. Steen, E. A. Vance, W. M. Woelmer, J. H. Wynne, and L. Zachmann. 2024.

Predicting spring phenology in deciduous broadleaf forests: NEON phenology forecasting community challenge. Agricultural and Forest Meteorology 345:109810.

Willson, A. M., H. Gallo, J. A. Peters, A. Abeyta, N. Bueno Watts, C. C. Carey, T. N. Moore, G. Smies, R. Q. Thomas, W. M. Woelmer, and J. S. McLachlan. 2023. Assessing opportunities and inequities in undergraduate ecological forecasting education. Ecology and Evolution 13:1–16.

Woelmer, W. M., L. M. Bradley, L. T. Haber, D. H. Klinges, A. S. L. Lewis, E. J. Mohr, C. L. Torrens, K. I. Wheeler, and A. M. Willson. 2021. Ten simple rules for training yourself in an emerging field. PLOS Computational Biology 17:e1009440.

Zhou, X., Y. Zhu, D. Hou, B. Fu, W. Li, H. Guan, E. Sinsky, W. Kolczynski, X. Xue, Y. Luo, J. Peng, B. Yang, V. Tallapragada, and P. Pegion. 2022. The Development of the NCEP Global Ensemble Forecast System Version 12. Weather and Forecasting 37:1069–1084.

Zwart, J. A., S. K. Oliver, W. D. Watkins, J. M. Sadler, A. P. Appling, H. R. Corson-Dosch, X. Jia, V. Kumar, and J. S. Read. 2023. Near-term forecasts of stream temperature using deep learning and data assimilation in support of management decisions. JAWRA Journal of the American Water Resources Association 59:317–337.

969 **Table 1.** Definitions of forecast uncertainty sources included in the submitted models, modified
970 from Dietze (2017), Thomas et al. (2020), and Lofton et al. (2023).

Source of uncertainty	**Definition**	**Example of how the uncertainty source could be quantified**
Process	Uncertainty from the inability of the model to replicate the dynamics of the forecasted state.	Calculating the error from the residuals of the model fit to historical data.
Parameter	Uncertainty in the parameter values of a fitted model.	Sampling from a distribution of parameter values and assigning different parameter values to each ensemble member.
Initial condition	Uncertainty in estimates of current conditions at the time of forecast generation (e.g., as a result of observation uncertainty, missing observations, and data assimilation).	Quantifying the spread in updated states following data assimilation or the previous day's forecast.
Driver	Uncertainty from driver data (e.g. future air temperature).	Using an ensemble of weather forecasts as drivers to the model.
Observation	Uncertainty from measurement error in the state being forecasted (difference between actual state and measured state).	Calculating the standard deviation of replicate water temperature observations.

971

Table 2. Representation of uncertainty within the best-performing water temperature (Tw) models (sorted in descending order) that had positive mean $CRPS_{skill}$ over the 1-30 day-ahead forecast horizon. See Table 1 for definitions of uncertainty types. For the comprehensive list of uncertainty sources for all submitted models and all variables, see Table S1.

Model	Source of uncertainty represented				
	Driver	**Parameter**	**Process**	**Initial conditions**	**Observation**
FLARE-GLM	x	x	x	x	x
FLARE-GLM-noDA	x	x	x	x	x
FLARE-GOTM	x	x	x	x	x
XGBoost	x		x		
Random Forest	x				
LER-Baselines MME	x	x	x	x	x
FLARE-LER MME	x	x	x	x	x
FLARE-GOTM-noDA	x	x	x	x	x
Prophet		x	x		
Lasso	x				

978 **Figure Captions**

979 **Figure 1**. Map of National Ecological Observatory Network (NEON) lake sites located across
980 the contiguous U.S., with map inset showing Alaska. Co-occurring sites are shown by the black
981 centroid and the coloured points are offset from this location. The points are labelled with their
982 four-character NEON site code: BARC (Barco Lake), CRAM (Crampton Lake), LIRO (Little
983 Rock Lake), PRLA (Prairie Lake), PRPO (Prairie Pothole Lake), SUGG (Suggs Lake), and
984 TOOK (Toolik Lake).

985 **Figure 2.** Mean relative skill ($CRPS_{skill}$, compared to day-of-year (DOY) baseline model) of
986 water temperature (Tw) and dissolved oxygen (DO) forecasts for the submitted models (averaged
987 across sites, submission dates, and 1-30 day-ahead horizons). Positive values indicate that a
988 submitted model performed better, on average, than the DOY baseline and negative values
989 indicate that the baseline performed better. Panel (a) shows the Tw models that outperformed the
990 DOY baseline as defined by $CRPS_{skill}$ and panel (b) shows all Tw models. Panel (c) shows
991 $CRPS_{skill}$ for DO models. The shading of the bars indicates the model structure; colour = model
992 class (empirical, machine learning (ML), multi-model ensemble (MME), process), and pattern =
993 inclusion of air temperature as a covariate. A second baseline model (Persistence), is shown in
994 grey (panels b,c) and models that outperformed the DOY baseline are highlighted by the grey
995 background shading.

996 **Figure 3.** Relative skill (a), mean standard deviation (b), and mean absolute bias (c) across the
997 30 day-ahead forecast horizon for the models that outperformed the day-of-year baseline for
998 water temperature. Relative skill was calculated as the difference in CRPS between the focal
999 model and the day-of-year baseline. The metrics in each panel were averaged across all sites and

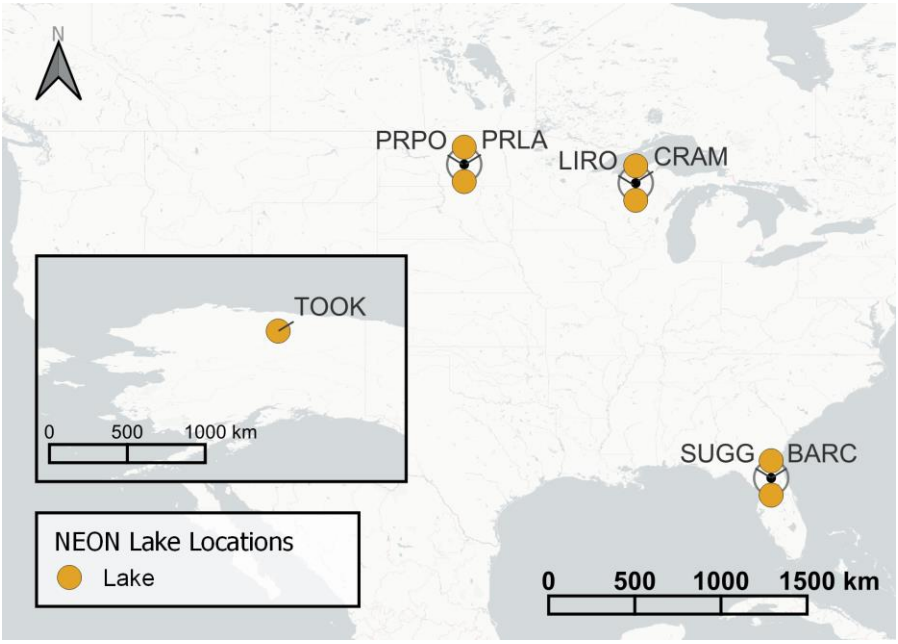
forecast submission dates. Models are listed in the legend in descending order of mean skill aggregated over the forecasting period.

Figure 4. Reliability plot (% of observations falling within the 95% and 80% confidence interval) for the water temperature models that out-performed the day-of-year baseline (a and b) and those that did not (grey lines, c and d). Perfectly confident forecasts would have an equal percentage of observations within the confidence interval as the percentage covered by the confidence interval. Values above the dashed line indicate that the forecast is underconfident (forecast precision is too wide) and values below the line indicate that the forecast is overconfident (forecast precision is too narrow). Values above the dotted threshold indicate that the forecast is underconfident (i.e., there are too many observations falling within the specified confidence interval) and values below the line indicate that the forecast is overconfident. Note the differences in scale between the panels a/b and c/d that show the 80% and 95% confidence intervals, respectively.

Figure 5. (a) Relative skill of water temperature forecasts compared to the baseline (day-of-year) for each site compared among model classes: empirical, machine learning (ML), multi-model ensemble (MME), and process-based. Positive values indicate the submitted model performed better, on average, than the baseline and negative values indicate that the baseline performed better. The *n* value indicates the number of models represented in each model class. (b) Mean relative skill for the top ten performing models among sites across the forecast horizon.

Figure 6. Difference in median monthly surface water temperature (depths < 1 m) between 2023 and historical observations (2015-2022) at the seven lake sites. Shaded regions show delta values that exceed 1 ^{o}C from median historical conditions. Not all lakes have historical observations for the full eight-year historical period or observations during all months.

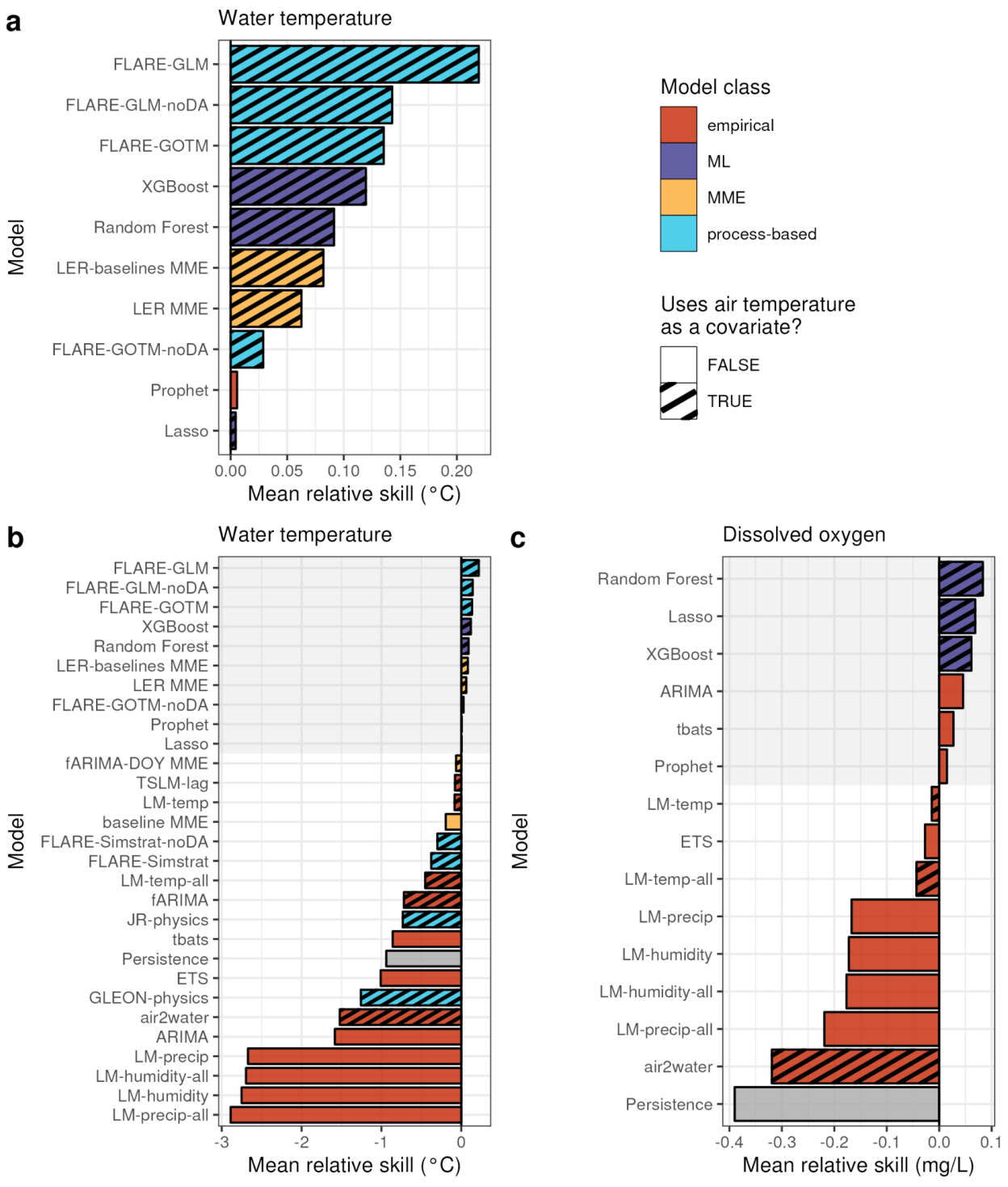
1023 Figure 1.



1024

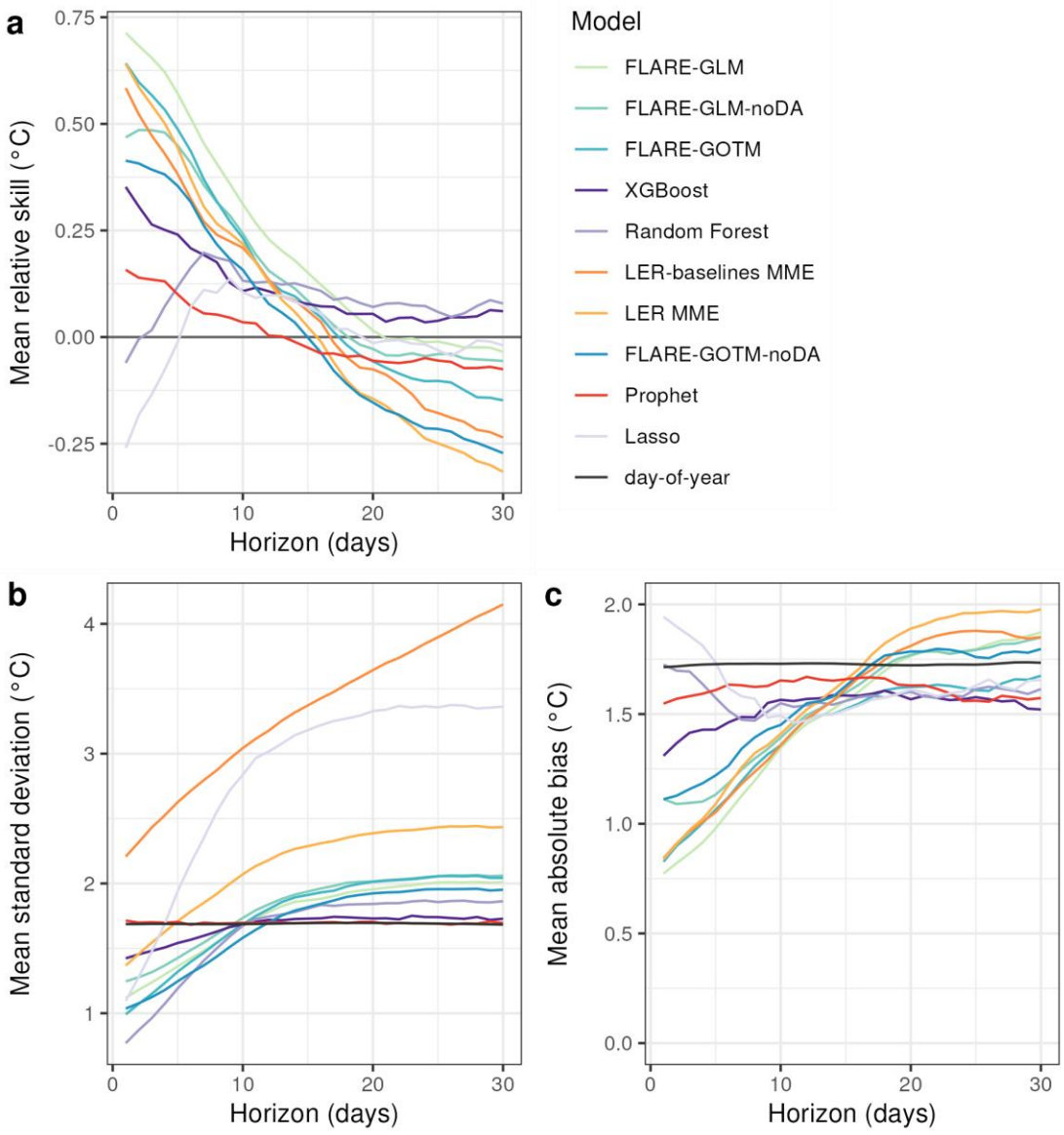
1025

1026 Figure 2.



1027

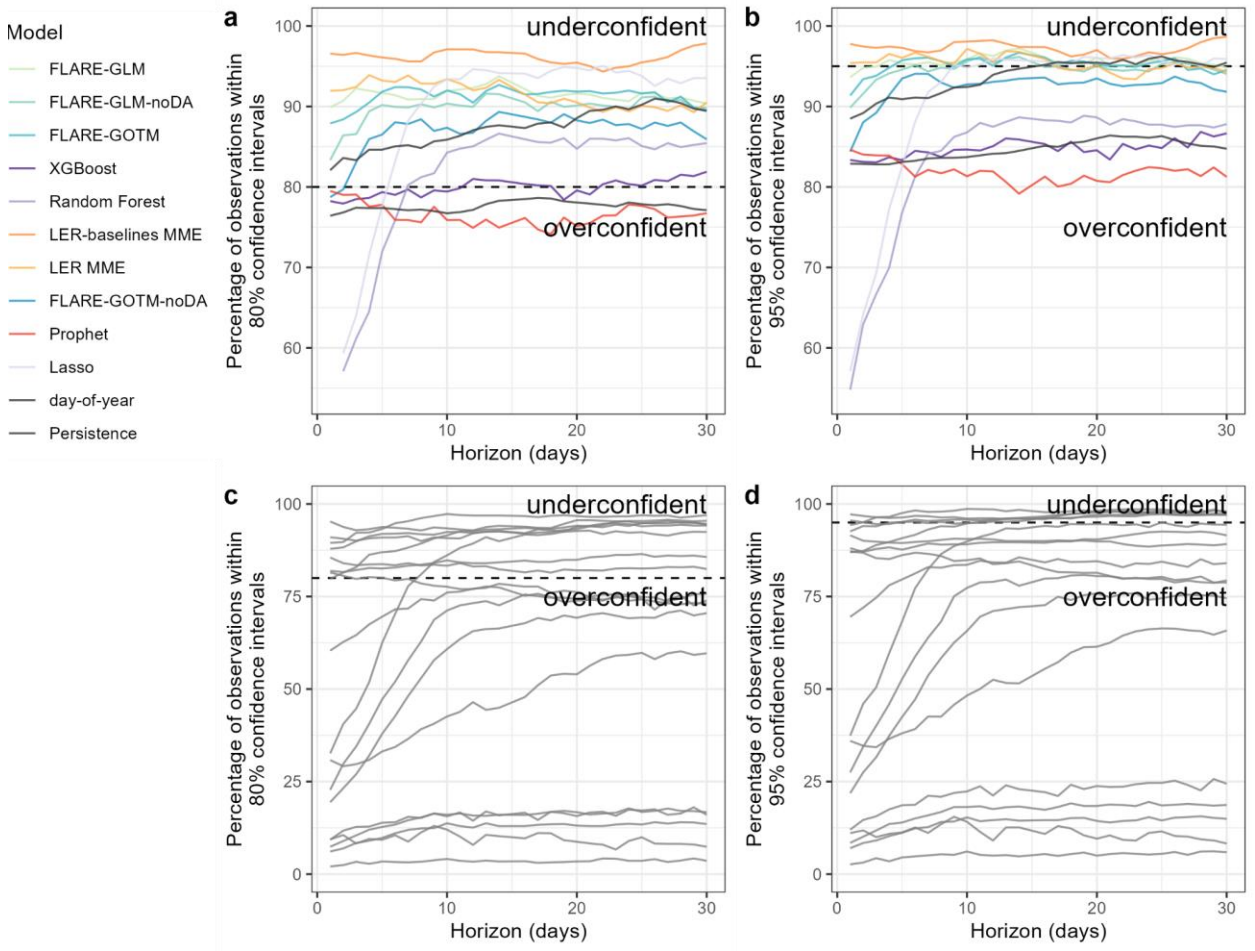
1028 Figure 3.



1029

1030

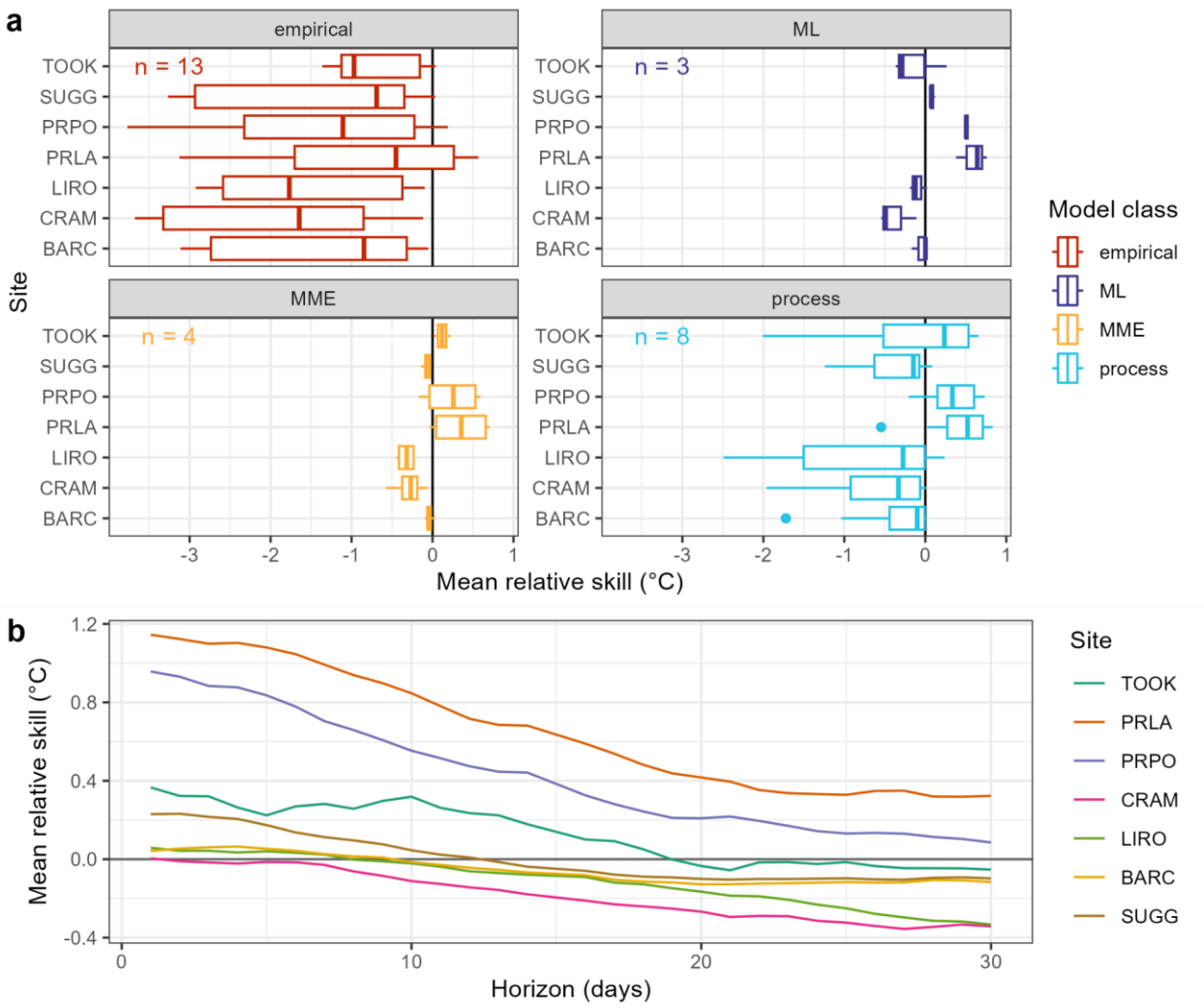
1031 Figure 4.



1032

1033

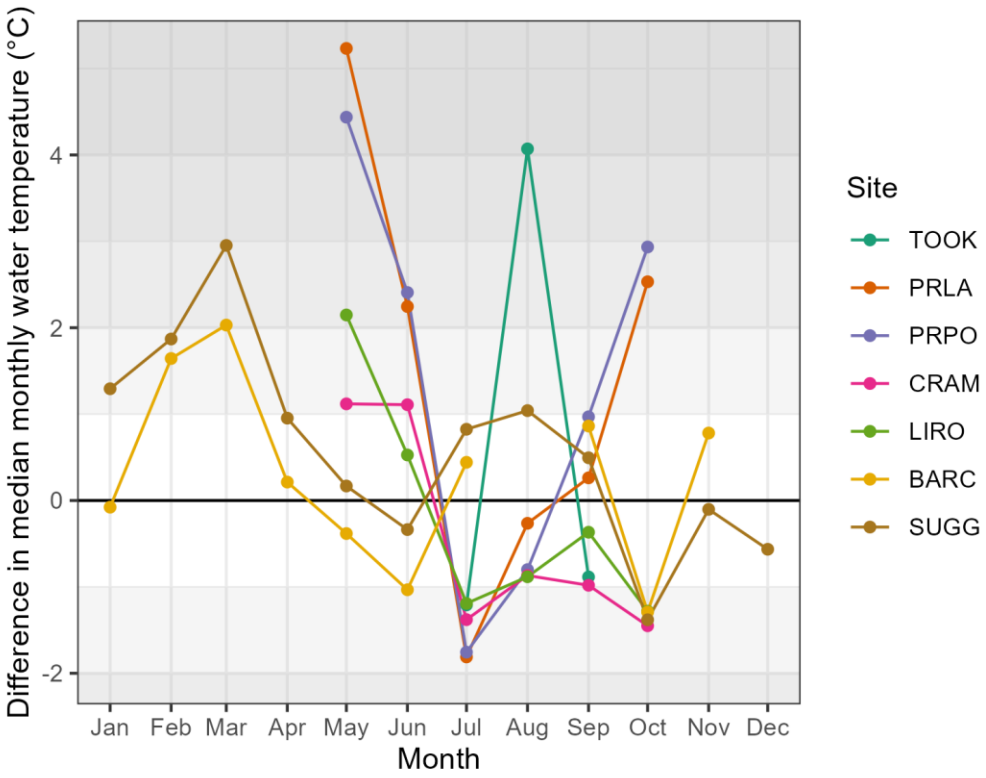
1034 Figure 5.



1035

1036

1037 Figure 6.



1038

Supplementary Information for *"What can we learn from 100,000 freshwater forecasts? A synthesis from the NEON Ecological Forecasting Challenge"*

Freya Olsson, Cayelan C. Carey, Carl Boettiger, Gregory Harrison, Robert Ladwig, Marcus F. Lapeyrolerie, Abigail S. L. Lewis, Mary E. Lofton, Felipe Montealegre-Mora, Joseph S. Rabaey, Caleb J. Robbins, Xiao Yang, R. Quinn Thomas

For submission to *Ecological Applications*

This supplement contains:

Figures: 4

Tables: 1

Texts: 1

Pages: 35

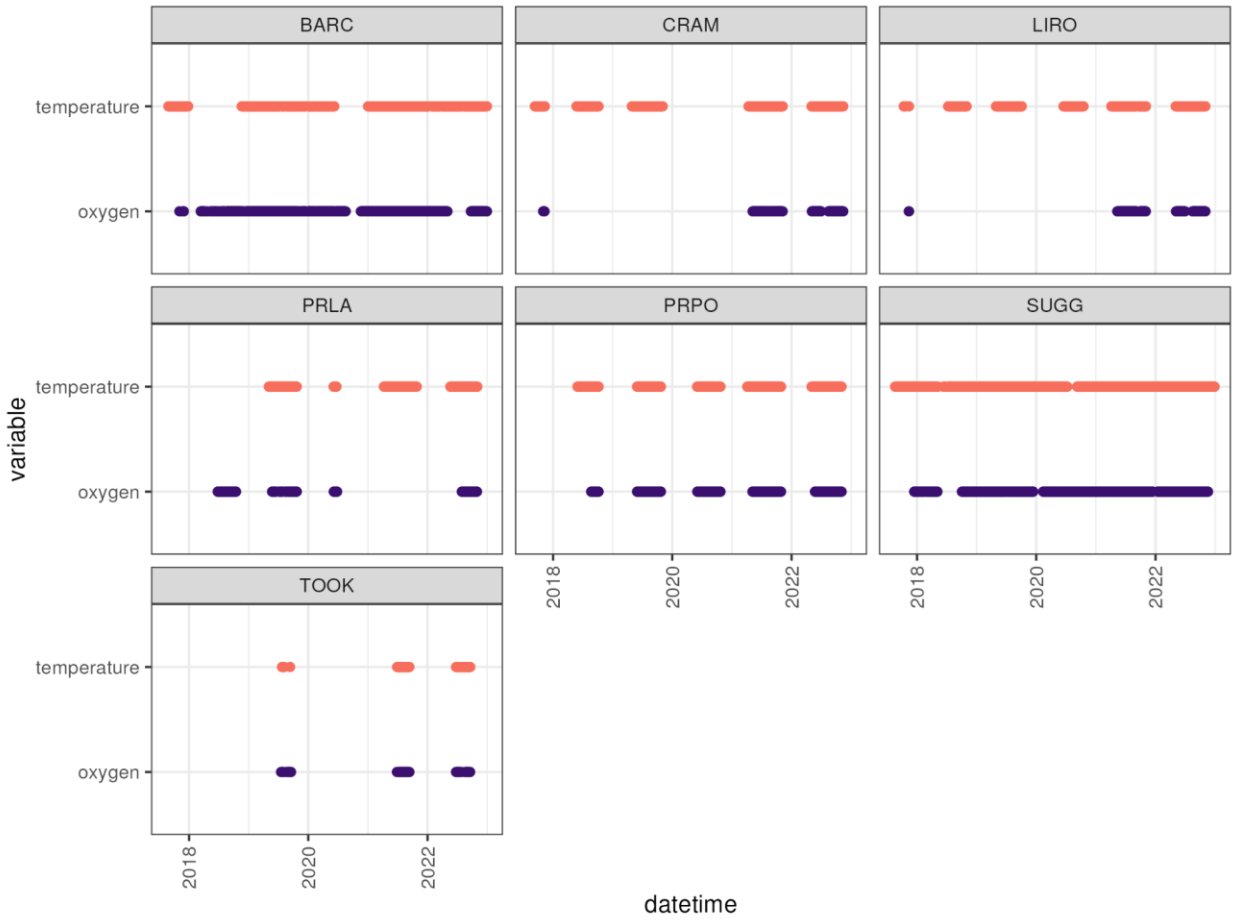


Figure S1. Time series of the water temperature and dissolved oxygen targets data available for each lake site and variable within the NEON Forecasting Challenge. Each panel shows data from a site denoted by the four-character NEON site code.

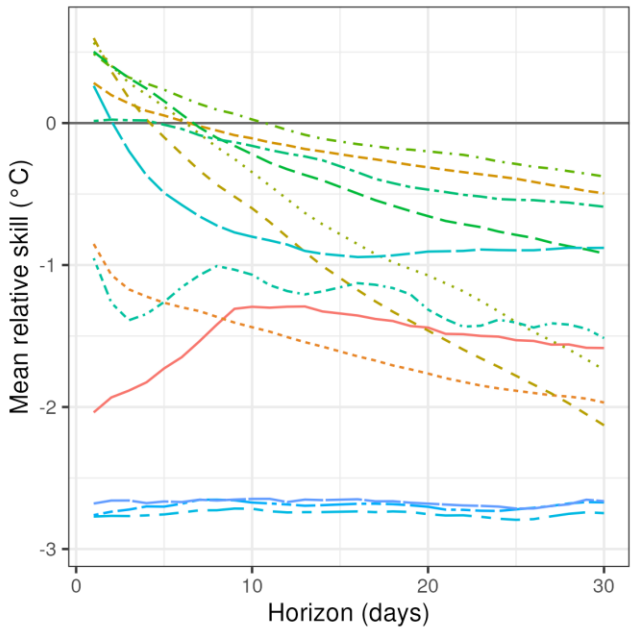


Figure S2. Mean relative skill of temperature (Tw) models that did not out-perform the day-of-year baseline (on average for all forecasts and sites) across the 30 day-ahead forecast horizon. Negative relative skill indicates that the baseline performed better and positive relative skill indicates the submitted model performed better.

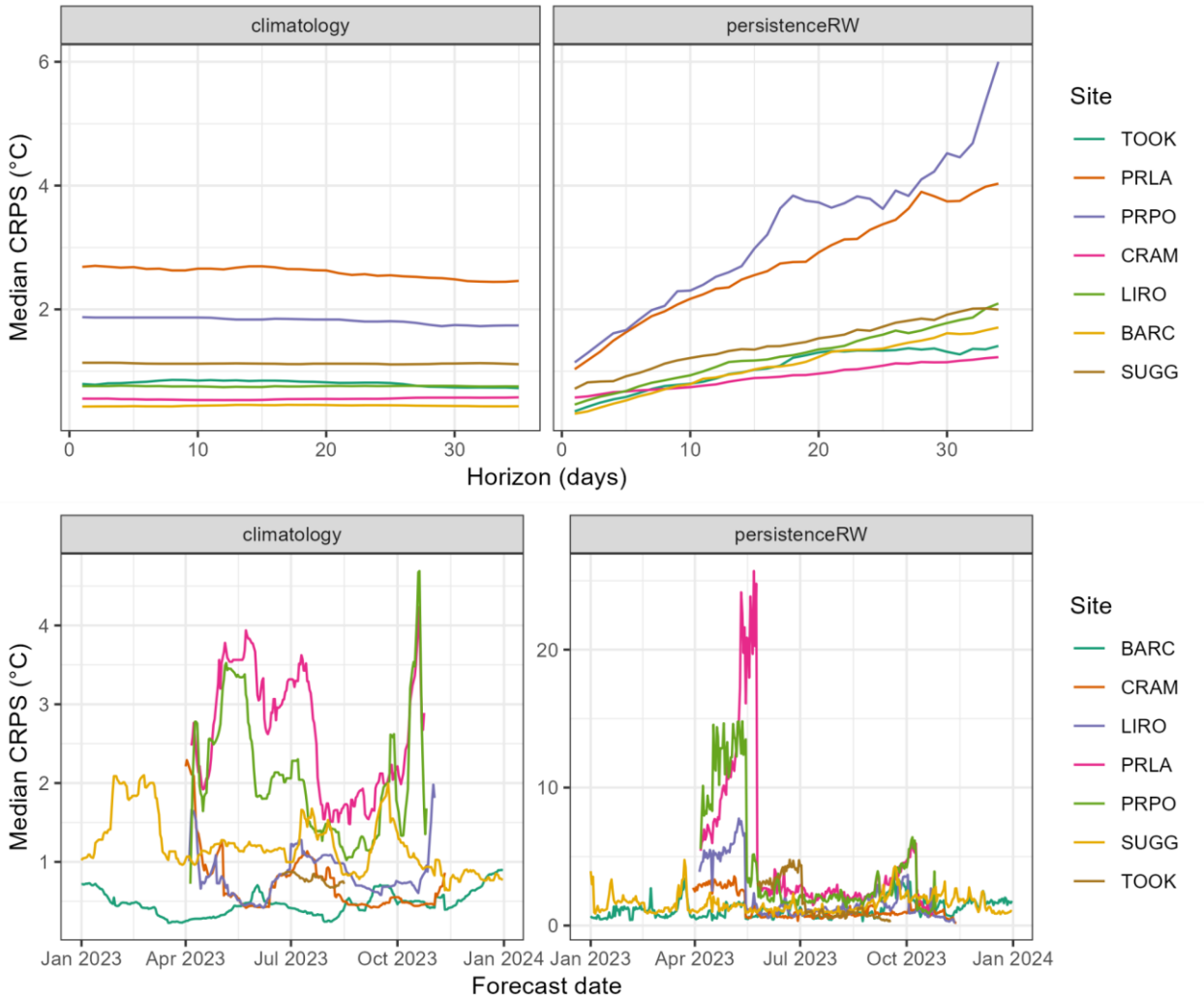


Figure S3. Median water temperature forecast performance (CRPS, °C) of the two baseline models (day-of-year/climatology, and persistence/persistenceRW) across the forecast horizon (A, B), and over time (median performance for each forecast date for all 1-30 days-ahead (C, D) at the seven lake sites.

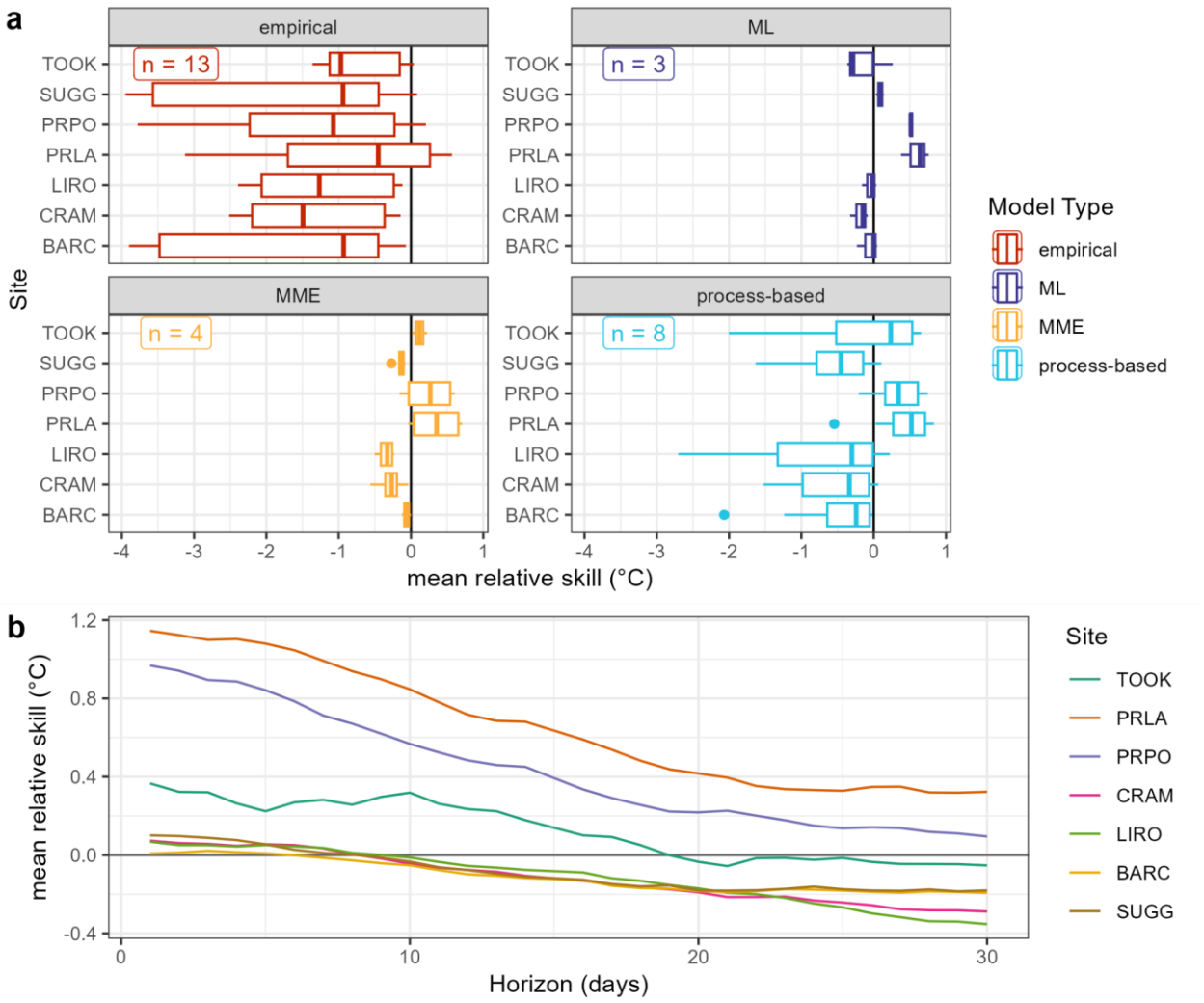


Figure S4. a) Relative skill of water temperature forecasts compared to the baseline (day-of-year) for each site compared among model classes (empirical, machine-learning (ML), multi-model ensembles (MME), and process). Positive values indicate the submitted model performed better, on average, than the baseline and negative values indicate that the baseline performed better. The *n* value indicates the number of models represented in each model class. b) Average relative skill for the top ten performing models among sites across the forecast horizon. Duration of forecasts was consistent among sites.

Table S1. A summary of model structure (type, covariates), sources of uncertainty, and use of historical data to produce initial conditions and update parameters for 30 forecast models submitted to the aquatics theme of the NEON Forecasting Challenge. A full description of the models can be found in Supplementary Text 1. Definitions for uncertainty types can be found in Table 2 in the main manuscript. *MME = multi-model ensemble.

Model Name	Model Type	Forecast Variables	Model Covariates	Includes Initial Conditions?	Is Model Dynamic?	Sources Of Uncertainty Represented	Updates Parameters?
Air2Water	Empirical	Temperature, Oxygen	Air Temperature	No	No	Driver	Yes
Baseline MME	MME	Temperature	-	Yes	Yes	Process	No
Prophet	Empirical	Temperature, Oxygen	-	No	Yes	Parameter, Process	Yes
Day-Of-Year	Null	Temperature, Oxygen	-	No	No		No
fARIMA	Empirical	Temperature	Air Temperature	Yes	Yes	Process, Driver	Yes
fARIMA-DOY MME*	MME	Temperature	Air Temperature	Yes	Yes	Process	Yes
LER MME*	MME	Temperature	Air Temperature, Air Pressure, Relative Humidity, Surface Downwelling Longwave, Surface Downwelling Shortwave, Precipitation, Eastward Wind, Northward Wind	Yes	Yes	Parameter, Process, Initial Conditions, Driver, Observation	Yes
LER-Baselines MME*	MME	Temperature	Air Temperature, Air Pressure, Relative Humidity, Surface Downwelling Longwave, Surface Downwelling Shortwave, Precipitation, Eastward Wind, Northward Wind	Yes	Yes	Parameter, Process, Initial Conditions, Driver, Observation	Yes

FLARE-GLM	Process	Temperature	Air Temperature, Air Pressure, Relative Humidity, Surface Downwelling Longwave, Surface Downwelling Shortwave, Precipitation, Eastward Wind, Northward Wind	Yes	Yes	Parameter, Process, Initial Conditions, Driver, Observation	Yes
FLARE-GLM- noDA	Process	Temperature	Air Temperature, Air Pressure, Relative Humidity, Surface Downwelling Longwave, Surface Downwelling Shortwave, Precipitation, Eastward Wind, Northward Wind	No	No	Parameter, Process, Initial Conditions, Driver, Observation	No
FLARE-GOTM	Process	Temperature	Air Temperature, Air Pressure, Relative Humidity, Surface Downwelling Longwave, Surface Downwelling Shortwave, Precipitation, Eastward Wind, Northward Wind	Yes	Yes	Parameter, Process, Initial Conditions, Driver, Observation	Yes
FLARE-GOTM-noDA	Process	Temperature	Air Temperature, Air Pressure, Relative Humidity, Surface Downwelling Longwave, Surface Downwelling Shortwave, Precipitation, Eastward Wind, Northward Wind	No	No	Parameter, Process, Initial Conditions, Driver, Observation	No

FLARE-Simstrat	Process	Temperature	Air Temperature, Air Pressure, Relative Humidity, Surface Downwelling Longwave, Surface Downwelling Shortwave, Precipitation, Eastward Wind, Northward Wind	Yes	Yes	Parameter, Process, Initial Conditions, Driver, Observation	Yes
FLARE-Simstrat-noDA	Process	Temperature	Air Temperature, Air Pressure, Relative Humidity, Surface Downwelling Longwave, Surface Downwelling Shortwave, Precipitation, Eastward Wind, Northward Wind	No	No	Parameter, Process, Initial Conditions, Driver, Observation	No
TSLM-Lag	Empirical	Temperature	Air Temperature	No	No	Process, Driver	Yes
JR-Physics	Process	Temperature	Air Temperature	Yes	Yes	Driver	No
GLEON-Physics	Process	Temperature	Air Temperature, Relative Humidity, Surface Downwelling Shortwave, Eastward Wind, Northward Wind	Yes	Yes	Process	No
Persistence	Null	Temperature, Oxygen	-	Yes	Yes	Process	No
ARIMA	Empirical	Temperature, Oxygen	-	Yes	Yes	Process	Yes
ETS	Empirical	Temperature, Oxygen	-	Yes	Yes	Process	Yes
LM-Humidity	Empirical	Temperature, Oxygen	Relative Humidity	No	No	Driver	Yes
LM-Humidity-All	Empirical	Temperature, Oxygen	Relative Humidity	No	No	Driver	Yes

Lasso	ML	Temperature, Oxygen	Air Temperature, Air Pressure, Relative Humidity, Surface Downwelling Longwave, Surface Downwelling Shortwave, Precipitation Flux, Northward Wind, Eastward Wind	No	No	Driver	No
LM-Precip	Empirical	Temperature, Oxygen	Precipitation	No	No	Driver	Yes
LM-Precip-All	Empirical	Temperature, Oxygen	Precipitation	No	No	Driver	Yes
Random Forest	ML	Temperature, Oxygen	Air Temperature, Air Pressure, Relative Humidity, Surface Downwelling Longwave, Surface Downwelling Shortwave, Precipitation, Eastward Wind, Northward Wind	No	No	Driver	No
TBATS	Empirical	Temperature, Oxygen	-	Yes	Yes	Process	Yes
LM-Temp	Empirical	Temperature, Oxygen	Air Temperature	No	No	Driver	Yes
LM-Temp-All	Empirical	Temperature, Oxygen	Air Temperature	No	No	Driver	Yes
XGBoost	ML	Temperature, Oxygen	Air Temperature, Surface Downwelling Shortwave, Relative Humidity	No	No	Process, Driver	Yes

Text S1. This supplementary text contains descriptions of the submitted models included in this paper. The model descriptions are provided by the forecast teams and include a description of the model's structure and general forecasting methodology. Links to the automated repositories are provided.

air2water (air2waterSat_2)

The air2water model is a linear model fit using the function lm() in R and uses air temperature as a covariate. The model fits water temperature (Tw) as a function of air temperature (Ta) and generates a forecast using forecasted water temperatures, following:

$$Tw = Ta * \beta_0 + \beta_1$$

where β_0 is a slope term and β_1 is an intercept. The uncertainty in drivers was obtained by using the 31 ensemble members from the NOAA GEFS forecast.

From these forecasted water temperatures, the dissolved oxygen concentration was estimated assuming 100% saturation of oxygen within the water (based on the temperature and elevation-dependent state calculation). To estimate the concentration of dissolved oxygen at saturation, the Eq.Ox.conc() in the *rMR* R package was used.

This model was used to forecast water temperature and dissolved oxygen concentration at the seven lake sites, with the model fitted separately for each site.

Team members: This model was generated as an example model by EFI-NEON Challenge Organisers.

Code repository: https://github.com/rqthomas/neon4cast-example/blob/main/forecast_model.R

Baseline MME (Baseline_ensemble)

The Baseline MME is a multi-model ensemble (MME) comprised of the two baseline models (day-of-year, persistence) submitted by Challenge organisers. To generate the MME, an ensemble forecast was generated by sampling from the submitted models (either from the ensemble members in the case of the persistence, or from the distribution for the day-of-year forecasts). The forecast included 200

ensemble members, represented equally across the 2 individual models (100 per forecast). The steps to generate the MME were:

1. Access submitted forecasts for the site and variable of interest (lake temperatures only) from the submissions S3 bucket.
2. Subset by individual model ID.
3. If the forecast is a distributional forecast: sample from the distribution using the forecasted mean and standard deviation to generate a sample of n = 100.
4. If the forecast is an ensemble forecast: subsample the existing individual forecast ensemble members to generate n = 100 ensemble members. The parameter numbers (ensemble members) were consistent across the forecast horizon.

Only sites with all individual forecasts present were submitted (a site, variable, and date had to be represented by both models).

This model was used to forecast water temperature in the lake sites (BARC, CRAM, LIRO, PRLA, PRPO, SUGG, TOOK). See information about the individual forecast models for information of forecast uncertainty representation in each of the forecasts.

Team members: Freya Olsson

Code repository: https://github.com/OlssonF/NEON-simple-baselines/blob/main/Models/baseline_ensemble.R

Prophet model (cb_prophet)

The Prophet model is an empirical model, specifically a non-linear regression model that includes seasonality effects (Taylor & Letham, 2018). The model relies on Bayesian estimation with an additive white noise error term:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t,$$

where g is a piecewise linear 'growth' term (with changepoints estimated by the algorithm), s is a seasonal effect (Fourier term), h is the effect of 'holidays', and epsilon is the white noise (error term). The

model does not include any covariate. We use the implementation of the Prophet model provided in the *darts* Python package (Herzen et al., 2022). See https://github.com/unit8co/darts

This model was used to forecast water temperature and dissolved oxygen concentration in the seven lake sites, with the model fitted separately for each site.

Code repository: https://github.com/cboettig/forecasts-darts-framework

Team members: Carl Boettiger, Marcus Francois Lapeyrolerie, Felipe Montealegre-Mora

Day-of-year (climatology)

The day-of-year (climatology) is a baseline model that assumes that forecasted conditions will be the same as the average of historical observations for that day-of-year (DOY). For each variable/site combination, the model calculates the mean (μ) and the standard deviation (σ) of the historical observations for each DOY. We assume that σ is consistent across the forecast horizon and so the median DOY σ is calculated for each new forecast (which can change between forecast dates but not across a forecast horizon).

Because of differences in sensor deployment (e.g., some lake sensors are removed in winter), not all DOYs have observations. Missing DOY means are filled using a linear interpolation, as long as at least two DOYs have values during the forecast period.

For the year 2023, the forecasts for each DOY do not change among forecast dates as no new data were collected during the forecast period (1 January 2023 - 31 December 2023) that would contribute to updated means or standard deviations.

This model was used to forecast water temperature and dissolved oxygen concentration in the seven lake sites, with the model fitted separately for each site.

Team members: this model was generated as a baseline model by EFI-NEON Challenge Organisers

Code repository: https://github.com/eco4cast/neon4cast-ci/blob/main/baseline_models/models/aquatics_climatology.R

fARIMA (fARIMA)

The fARIMA is an empirical model that fits an ARIMA model using the *fable* R package (O'Hara-Wild, Hyndman, & Wang 2023) as a function of a linear model with air temperature. The default ARIMA() function automatically chooses the best ARIMA model for the time-series, using a step-wise procedure.

The process uncertainty is generated from the standard deviation in the residuals of the fitted model. We could not assume a normal distribution in residuals and so opted to generate an ensemble forecast using a bootstrap approach within the generate() function from *fable*. In addition, we used the 31 ensemble members from the NOAA GEFS as driver uncertainty. For each NOAA ensemble member, an ensemble forecast with six ensemble members was generated using the generate() function resulting in a total of 31 x 6 = 186 ensemble members per forecast.

Not all sites have observations for all days due to differences in maintenance (e.g., some lake sites have sensors removed in winter). Therefore, to account for the difference in the time since last observation, the forecast was started at the day after the last observation, and the horizon modified to cover up to 30 days into the future from the forecast date. During this 'catch-up' period, the pseudo-observation of air temperature used in model training was used to generate water temperature rather than forecasted air temperature.

This model was used to forecast water temperature in the seven lake sites, with the model fitted separately for each site.

Team members: Freya Olsson, R. Quinn Thomas

Code repository: https://github.com/OlssonF/NEON-simple-baselines/blob/main/Models/ARIMA_model.R

fARIMA-DOY MME (fARIMA_clim_ensemble)

The fAMIRA-DOY MME is a multi-model ensemble (MME) composed of two empirical models: an ARIMA model (fARIMA) and day-of-year model. To generate the MME, an ensemble forecast was generated by sampling from the submitted models' ensemble members. The forecast included 200 ensemble members, represented equally across the two individual models (n=100). The steps to generate the MME were:

1. Access submitted forecasts for the site and variable of interest (lake temperatures only) from the submissions S3 bucket.
2. Subset by individual model ID.
3. Subsample the existing individual forecast ensemble members to generate 100 ensemble members. The parameter numbers (ensemble members) were consistent across the forecast horizon.
4. Only sites with all individual forecasts present were submitted (a site, variable and date had to be represented by both models).

This model was used to forecast water temperature in the seven lake sites, with the model fitted separately for each site. See information about the individual forecast models for information of forecast uncertainty representation in each of the forecasts.

Team members: Freya Olsson

Code repository: https://github.com/OlssonF/NEON-simple-baselines/blob/main/Models/fARIMA_clim_ensemble.R

LER MME (flare_ler)

The LER MME is a multi-model ensemble (MME) derived from the three process models from FLARE (FLARE-GLM, FLARE-GOTM, and FLARE-Simstrat). To generate the MME, an ensemble forecast was generated by sampling from the submitted models' ensemble members. The forecast included 198 ensemble members, represented equally across the 3 individual models (n=66). The steps to generate the MME were:

1. Access submitted forecasts for the site and variable of interest (lake temperatures only) from the submissions S3 bucket.

2. Subset by individual model ID.

3. Subsample the existing individual forecast ensemble members to generate 66 ensemble members. The parameter numbers (ensemble members) were consistent across the forecast horizon.

4. Only sites with all individual forecasts present were submitted (a site, variable and date had to be represented by all 3 models).

This model was used to forecast water temperature in six lake sites (BARC, CRAM, LIRO, PRLA, PRPO, SUGG), but not TOOK, with the model fitted separately for each site. See information about the individual forecast models for information of forecast uncertainty representation in each of the forecasts.

Team members: Freya Olsson

Code repository: https://github.com/OlssonF/NEON-simple-baselines/blob/main/Models/flare_ler_ensemble.R

LER-baselines MME (flare_ler_baselines)

The LER-baselines model is a multi-model ensemble (MME) comprised of the three process models from FLARE (FLARE-GLM, FLARE-GOTM, and FLARE-Simstrat) and the two baseline models (day-of-year, persistence), submitted by Challenge organisers. To generate the MME, an ensemble forecast was generated by sampling from the submitted model's ensemble members (either from an ensemble forecast in the case of the FLARE models and persistence, or from the distribution for the day-of-year forecasts). The forecast included 200 ensemble members, represented equally across the 5 individual models (40 per forecast). The steps to generate the MME were:

1. Access submitted forecasts for the site and variable of interest (lake temperatures only) from the submissions S3 bucket.

2. Subset by individual model ID.

3. If the forecast is a distributional forecast: sample from the distribution using the forecasted mean and standard deviation to generate a sample of n = 40.

4. If the forecast is an ensemble forecast: subsample the existing individual forecast ensemble members to generate n = 40 ensemble members. The parameter numbers (ensemble members) were consistent across the forecast horizon.

5. Only sites with all individual forecast present were submitted (a site, variable, and date had to be represented by all 5 models).

This model was used to forecast water temperature in six of the lake sites (BARC, CRAM, LIRO, PRLA, PRPO, SUGG), but not TOOK. See information about the individual forecast models for information of forecast uncertainty representation in each of the forecasts.

Team members: Freya Olsson

Code repository: https://github.com/OlssonF/NEON-simple-baselines/blob/main/Models/ler_baselines_ensemble.R

FLARE-GLM (flareGLM)

The FLARE-GLM is a forecasting framework that integrates the General Lake Model hydrodynamic process model (GLM; Hipsey et al., 2019) and data assimilation algorithm to generate ensemble forecasts of lake water temperature. FLARE's ensemble-based forecasting algorithm generates forecasts using GLM that quantifies the uncertainty from driver data (weather forecasts from NOAA's Global Ensemble Forecasting System; Hamill et al., 2022), initial conditions, model process, and model parameters and then samples from these sources of uncertainty to generate probability distributions for water temperature at multiple lake or reservoir depths (see Thomas et al., 2020).

Daily forecasts were generated for the lake sites using the following steps: Step 1) access the FLARE-GLM forecasts from the day before (or, in the case of the first forecast, following a 60 day spin-up); Step 2) use this prediction to initialise a GLM run that starts 5 days ago and runs to current day; Step 3) use the ensemble Kalman filter (Evensen, 2003) to assimilate new observations collected over the past

5 days to update GLM's states and parameters; and Step 4) use the updated states and parameters as initial conditions for a 1- to 30 day-ahead forecast that starts today. Each forecast includes 256 ensemble members that quantify the uncertainty from driver data (weather forecasts), initial conditions, model process, and model parameters.

Driver uncertainty: GLM requires the following weather covariates obtained from NOAA GEFS: air temperature, air pressure, relative humidity, wind speed (calculated from the north and east wind speeds), precipitation, and incoming shortwave and incoming longwave radiation. The water balance method was set to include no inflows or outflows and maintain a water level. Bathymetry data for the lakes were obtained from NEON. Uncertainty from drivers was generated based on the 31 ensemble members from NOAA GEFS.

Initial conditions uncertainty: Initial conditions uncertainty was based on the spread of model states on Day 0 of the forecast that was set by spread in the 256 ensemble members following data assimilation on Day 0.

Model process uncertainty: Process uncertainty was generated by adding random noise to each ensemble, drawing from a normal distribution with a standard deviation of 0.75 °C (after Thomas et al., 2020).

Model parameter: parameter uncertainty was generated using a unique parameter value assigned to each of the 256 ensemble members that was determined through data assimilation. The parameters that are tuned in the data assimilation algorithm are specific to the hydrodynamic model. In total, two parameters were tuned in the data assimilation process: lw_factor (longwave radiation scaling factor), and sed_temp_mean (annual mean sediment temperature, °C).

Forecasts of daily surface water temperature were generated from the profiles output from FLARE-GLM by averaging temperatures forecasted in the top 1 m of the water column as a "surface" forecast. FLARE-GLM outputs forecasts for 00:00:00 and this is given as the daily forecast to the Challenge.

This model was used to forecast water temperature at the seven lake sites with the model parameters calibrated separately for each site. Additional information about FLARE configuration can be found in Thomas et al. (2020) and Thomas et al. (2023).

Team members: Freya Olsson, R. Quinn Thomas, Cayelan C. Carey

Code repository: https://github.com/FLARE-forecast/NEON-forecast-code/workflows/default

FLARE-GLM-noDA (flareGLM_noDA)

FLARE-GLM-noDA uses the same configuration as FLARE-GLM with the exception of the data assimilation (DA) algorithm. Within the noDA configuration, model states and parameters were not updated prior to forecast generation. Model parameters were calibrated before using observations of water temperatures and then brought 'online' to generate real-time forecasts using forecast drivers (NOAA weather data). Parameter uncertainty was calculated (as in FLARE-GLM), but the distributions were not updated between forecasts. The parameters calibrated were lw_factor (longwave radiation scaling factor), and sed_temp_mean (annual mean sediment temperature, °C).

Forecasts of daily surface water temperature were generated from the profiles output from FLARE-GLM-noDA by averaging temperatures forecasted in the top 1 m of the water column as a "surface" forecast. FLARE-GLM-noDA outputs forecasts for 00:00:00 and this is given as the daily forecast to the Challenge.

This model was used to forecast water temperature at the seven lake sites, with the model parameters calibrated separately for each site. Additional information about FLARE configuration can be found in Thomas et al., (2020) and Thomas et al. (2023).

Team members: Freya Olsson, R. Quinn Thomas

Code repository: https://github.com/FLARE-forecast/NEON-forecast-code/workflows/default

FLARE-GOTM (flareGOTM)

FLARE-GOTM uses the same principles and overarching framework as FLARE-GLM, with the hydrodynamic model replaced with the General Ocean Turbulence Model (GOTM). GOTM is a 1-D hydrodynamic turbulence model (Umlauf et al., 2005) that estimates water column temperatures. The integration of FLARE and GOTM was achieved using the LakeEnsemblR R package (Moore et al., 2021). Sources of uncertainty remain the same and are generated using equivalent methods. The parameters that were tuned in the data assimilation algorithm were specific to the hydrodynamic model and in the case of GOTM are swr_scale (short-wave radiation scaling factor) and/or wind_scale (wind speed, u10, scaling factor), depending on the site's sensitivity. See FLARE-GLM for a full description of the sources of uncertainty and the forecast generation method.

Forecasts of daily surface water temperature were generated from the profiles output from FLARE-GOTM by averaging temperatures forecasted in the top 1 m of the water column as a "surface" forecast. FLARE-GLM outputs forecasts for 00:00:00 and this is given as the daily forecast to the Challenge.

This model was used to forecast water temperature in 6 of the lake sites (BARC, CRAM, LIRO, PRLA, PRPO, SUGG), but not TOOK, with the model parameters calibrated separately for each site. Additional information about FLARE configuration can be found in Thomas et al., (2020) and Thomas et al. (2023).

Team members: Freya Olsson, R. Quinn Thomas

Code repository: https://github.com/FLARE-forecast/NEON-forecast-code/workflows/ler

FLARE-GOTM-noDA (flareGOTM_noDA)

FLARE-GOTM-noDA uses the same configuration as FLARE-GLM with the exception of the data assimilation (DA) algorithm. Within the noDA configuration, model states and parameters are not updated prior to forecast generation. Model parameters were calibrated before using observations of water temperatures and then brought 'online' to generate real-time forecasts using forecast drivers (NOAA

weather data). Parameter uncertainty was calculated (as in FLARE-GOTM) but the distributions were not updated between forecasts. The parameters calibrated were swr_scale (short-wave radiation scaling factor), and/or wind_scale (wind speed, u10, scaling factor).

Forecasts of daily surface water temperature were generated from the profiles output from FLARE-GOTM-noDA by averaging temperatures forecasted in the top 1 m of the water column as a "surface" forecast. FLARE-GOTM-noDA outputs forecasts for 00:00:00, which is submitted as the daily forecast to the Challenge.

This model was used to forecast water temperature in six of the lake sites (BARC, CRAM, LIRO, PRLA, PRPO, SUGG), but not TOOK, with the model parameters calibrated separately for each site. Additional information about FLARE configuration can be found in Thomas et al. (2020) and Thomas et al. (2023).

Team members: Freya Olsson, R. Quinn Thomas

Code repository: https://github.com/FLARE-forecast/NEON-forecast-code/workflows/ler

FLARE-Simstrat (flareSimstrat)

FLARE-Simstrat uses the same principles and overarching framework as FLARE-GLM with the hydrodynamic model replaced with Simstrat. Simstrat is a 1-D hydrodynamic turbulence model (Goudsmit et al., 2002) that estimates water column temperatures. The integration of FLARE and Simstrat was achieved using the LakeEnsemblR R package (Moore et al., 2021). Sources of uncertainty remain the same and are generated using equivalent methods. The parameters that are tuned in the data assimilation algorithm are specific to the hydrodynamic model and in the case of Simstrat were p_sw_water (short-wave radiation scaling factor) and/or f_wind (wind speed scaling factor), depending on the site's sensitivity. See FLARE-GLM for a full description of the sources of uncertainty and the forecast generation method.

Forecasts of daily surface water temperature were generated from the profiles output from FLARE-Simstrat by averaging temperatures forecasted in the top 1 m of the water column as a "surface"

forecast. FLARE-Simstrat outputs forecasts for 00:00:00, which is submitted as the daily forecast for the Challenge.

This model was used to forecast water temperature at the seven lakes, with the model parameters calibrated separately for each site. Additional information about FLARE configuration can be found in Thomas et al. (2020) and Thomas et al. (2023).

Team members: Freya Olsson, R. Quinn Thomas

Code repository: https://github.com/FLARE-forecast/NEON-forecast-code/workflows/ler

FLARE-Simstrat-noDA (flareSimstrat_noDA)

FLARE-Simstrat-noDA uses the same configuration as FLARE-Simstrat with the exception of the data assimilation (DA) algorithm. Within the noDA configuration, model states and parameters were not updated prior to forecast generation. Model parameters were calibrated before using observations of water temperatures and then brought 'online' to generate real-time forecasts using forecast drivers (NOAA weather data). Parameter uncertainty was calculated (as in FLARE-Simstrat) but the distributions were not updated between forecasts. The parameters calibrated were p_sw_water (incoming short-wave radiation scaling factor), and/or f_wind (wind speed scaling factor).

Forecasts of daily surface water temperature were generated from the profiles output from FLARE-Simstrat-noDA by averaging temperatures forecasted in the top 1 m of the water column as a "surface" forecast. FLARE-Simstrat-noDA outputs forecasts for 00:00:00 and this is given as the daily forecast to the Challenge.

This model was used to forecast water temperature at the seven lake sites, with the model parameters calibrated separately for each site. Additional information about FLARE configuration can be found in Thomas et al. (2020) and Thomas et al. (2023).

Team members: Freya Olsson, R. Quinn Thomas

Code repository: https://github.com/FLARE-forecast/NEON-forecast-code/workflows/ler

TSLM-lag (fTSLM_lag)

This is a simple time series linear model in which water temperature is a function of air temperature of that day and the previous day's air temperature. The TSLM was fit using the TSLM() function from the *fable* R package (O'Hara-Wild M, Hyndman R, Wang E, 2023).

The process uncertainty is generated from the standard deviation in the residuals of the fitted model. We could not assume a normal distribution in residuals and so opted to generate an ensemble forecast using a bootstrap approach within the generate() function from *fable*. In addition, we used the 31 ensemble members from the NOAA GEFS as driver uncertainty. For each NOAA ensemble member, an ensemble forecast with six ensemble members was generated using the generate() function resulting in a total of 31 x 6 = 186 ensemble members per forecast.

Not all sites have observations for all days due to differences in maintenance (e.g., some lake sites have sensors removed in winter). Therefore, to account for the difference in the time since last observation, the forecast was started at the day after the last observation, and the horizon modified to cover up to 30 days into the future from the forecast date. During this 'catch-up' period the pseudo-observation of air temperature, used in model training, was to generate water temperature rather than forecasted air temperature.

This model was used to forecast water temperature at the seven lake sites, with the model fitted separately for each site.

Team members: Freya Olsson, R. Quinn Thomas

Code repository: https://github.com/OlssonF/NEON-simple-baselines/blob/main/Models/TSLM_lags.R

JR-physics (GLEON_JRabaey_temp_physics)

The JR-physics model is a simple process model based on the assumption that surface water temperature should trend towards equilibration with air temperature with a lag factor.

Initial conditions for the model were set using the most recently available temperature data for each site. Forecasted water temperature was calculated as:

$$Tw_{t+1} = Tw_t + l(Ta_{t+1} - Tw_t)$$

where *Tw* is surface water temperature, *Ta* is forecasted air temperature, and *l* is the air-water equilibration lag factor. *l* was set to 0.2 for all sites.

Each forecast is generated using the 31 ensemble air temperature forecasts from the NOAA GEFS weather forecast. The model is iteratively fit each day as new data are generated by NEON.

Driver uncertainty was included by using the 31 ensemble members from the NOAA GEFS weather forecast. No uncertainties from initial conditions or model process were included in the overall forecast uncertainty.

This model was used to forecast water temperature at the seven lake sites, with the model fitted separately for each site.

Team members: Joseph Rabaey

Code Repository: https://github.com/jrabaey/Neon4cast-JR-Physics

GLEON-physics (GLEON_physics)

A simple, process-based model was developed to replicate the water temperature dynamics of a surface water layer *sensu* Chapra (2008). The model focus was only on quantifying the impacts of atmosphere-water heat flux exchanges on the idealized near-surface water temperature dynamics:

$$\frac{\partial T}{\partial t} = \frac{1}{\Delta z}\frac{(Q + H)}{c_p \rho_w}$$

where *T* is water temperature, *t* is time (fixed time step of 3600 s), Δz is the thickness of the near-surface layer which is assumed to be 1 m, *Q* is the net heat flux, *H* is internal heat generation due to incoming short-wave radiation, c_p is the heat capacity of water, and ρ_w is water density. *Q* represents the amount of energy from short-wave radiation that is absorbed directly in the surface layer:

$$Q = (1 - \alpha)Q_{sw}$$

with α as a constant albedo of 0.1. The net heat flux H is the sum of four terms:

$$H = H_{lw} + H_{lwr} + H_v + H_c,$$

where the terms on the right-hand side represent incoming long-wave radiation, emitted long-wave radiation from the water, the latent heat flux, and the sensible heat flux, respectively.

The heat fluxes were derived using the formulations from Livingstone and Imboden (1989), Goudsmit et al. (2002), and Verburg and Antenucci (2010). Note that the latent and sensible heat fluxes were calculated by including the actual surface area of the respective lake. To replicate the heat flux dynamics, the model used mean forecasted air temperature, relative humidity, air pressure, short-wave radiation, and wind speed from NOAA GEFS. Air vapor pressure was quantified from air temperature and relative humidity. Cloud cover was calculated using the empirical formulation from Martin and McCutcheon (1998). Whenever water temperatures became less than the freezing point temperature of water (assumed to be 0 °C), water temperatures were set to 0 °C.

We approximated the water temperature of the next time step using an explicit Euler forward scheme, and also by including an error term on the right-hand side to account for stochastic fluctuations:

$$T_{t+1} = T_t + \frac{\Delta t}{\Delta z}\frac{(Q + H)}{c_p \rho_w} + N(\mu, \sigma)$$

where μ was set to 0 °C, and σ to 0.05 °C. For every prediction, we ran 100 model runs to quantify process uncertainty through the error term. No uncertainties from initial conditions, drivers, or parameter estimations were included in the overall forecast uncertainty.

This model was used to forecast water temperature at the seven lake sites, with the model fitted separately for each site.

Team members: Robert Ladwig, Xiao Yang

Code repository: https://github.com/robertladwig/NEON-simple-baselines/tree/main

Persistence (persistenceRW)

The persistence (persistenceRW; random walk) is a baseline model that assumes, on average, conditions over the forecast horizon will be the same as the last observation, with uncertainty driven by a random walk process.

$$y_{T+1} = y_T + e_{T+1}$$

where y_T is today's observation or forecast, e_{T+1} is random noise, and y_{T+1} is the next day's forecast. The uncertainty (e_{T+1}) in the persistence model forecasts was generated using a bootstrapping method with no assumption placed on the distribution of the forecast. We assumed that future uncertainty will be drawn from the same distribution of the residual error in the fit to historical data. We fit the model to historical observations, using the RW() (Random walk) function in the *fable* R package (version 0.3.2; O'Hara-Wild et al., 2022), and the model error or residual (e) was calculated between the model and observations. At each timestep, a value of e_{T+1} was drawn from the distribution of these historic error values for each ensemble member. Overall, 200 ensemble members were generated for each timestep using this method using the *fable* generate() function and a bootstrap value of 200 (number of ensemble members).

This model was used to forecast water temperature and dissolved oxygen concentration at the seven lake sites, with the model fitted separately for each site and variable.

Team members: this model was generated as a baseline model by EFI-NEON Challenge Organisers.

Code repository: https://github.com/eco4cast/neon4cast-ci/blob/main/baseline_models/models/aquatics_persistenceRW.R; https://github.com/eco4cast/neon4cast-ci/blob/main/baseline_models/R/fablePersistenceModelFunction.R

ARIMA (tg_arima)

The tg_arima model is an AutoRegressive Integrated Moving Average (ARIMA) model fit using the function auto.arima() from the *forecast* package in R (Hyndman et al. 2023; Hyndman et al., 2008). This is an empirical time series model with no covariates. The model is fit every day as new data are made available, and is fit separately for each site/variable combination. For sites/variables where all

observations were non-negative, we set the Box-Cox transformation parameter (lambda) in forecast::auto.arima() to “auto”, allowing a Box-Cox transformation to be automatically selected. Forecasts were generated based on the model fit using the forecast::forecast() function, and were submitted as normal distributions using the mean and standard deviation of the forecast output.

This model was used to forecast water temperature and dissolved oxygen concentration at the seven lake sites, with the model fitted separately for each site.

Team members: Abigail S.L. Lewis, Caleb J. Robbins

Code repository:

https://github.com/eco4cast/Forecast_submissions/blob/main/Generate_forecasts

ETS (tg_ets)

The tg_ets model is an Error, Trend, Seasonal (ETS) model fit using the function ets() from the *forecast* package in R (Hyndman et al. 2023; Hyndman et al., 2008). This is an empirical time series model with no covariates. The model is fit every day as new data are made available, and is fit separately for each site/variable combination. We interpolated all missing data in the time series for each site/variable combination using forecast::na.interp(). For sites/variables where all observations are non-negative, we set the Box-Cox transformation parameter (lambda) in *forecast::na.interp()* to “auto”, allowing a Box-Cox transformation to be automatically selected. Forecasts were generated based on the model fit using the forecast::forecast() function, and were submitted as normal distributions using the mean and standard deviation of the forecast output.

This model was used to forecast water temperature and dissolved oxygen concentration at the seven lake sites, with the model fitted separately for each site.

Team members: Abigail S.L. Lewis, Caleb J. Robbins

Code repository:

https://github.com/eco4cast/Forecast_submissions/blob/main/Generate_forecasts

TBATS (tg_tbats)

The tg_tbats model is a TBATS (Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components) model fit using the function tbats() from the *forecast* package in R (Hyndman et al. 2023; Hyndman et al., 2008). This is an empirical time series model with no covariates. The model is fit every day as new data are made available, and is fit separately for each site/variable combination. We interpolated all missing data in the time series for each site/variable combination using forecast::na.interp(). For sites/variables where all observations are non-negative, we set the Box-Cox transformation parameter (lambda) in *forecast::na.interp()* to "auto", allowing a Box-Cox transformation to be automatically selected. Forecasts were generated based on the model fit using the forecast::forecast() function, and were submitted as normal distributions using the mean and standard deviation of the forecast output.

This model was used to forecast water temperature and dissolved oxygen concentration at the seven lake sites, with the model fitted separately for each site.

Team members: Abigail S.L. Lewis, Caleb J. Robbins

Code repository:

https://github.com/eco4cast/Forecast_submissions/blob/main/Generate_forecasts

LM-humidity (tg_humidity_lm)

The tg_humidity_lm model is a linear model fit using the function lm() in R. This is a very simple model with only one covariate: relative humidity. The model is fit every day as new data are made available, and is fit separately for each site/variable combination.

Driver uncertainty was included by using the 31 ensemble members from the NOAA GEFS weather forecast. No uncertainties from initial conditions or model process were included in the overall forecast uncertainty.

This model was used to forecast water temperature and dissolved oxygen concentration at the seven lake sites, with the model fitted separately for each site.

Team members: Abigail S.L. Lewis, Caleb J. Robbins

Code repository:

https://github.com/eco4cast/Forecast_submissions/blob/main/Generate_forecasts

LM-humidity-all (tg_humidity_lm_all_sites)

The tg_humidity_lm_all_sites model is a linear model fit using the function lm() in R. This is a very simple model with only one covariate: relative humidity. The model is fit every day as new data are made available, and is fit across all sites, using site ID as a factor in the regression.

Driver uncertainty was included by using the 31 ensemble members from the NOAA GEFS weather forecast. No uncertainties from initial conditions or model process were included in the overall forecast uncertainty.

This model was used to forecast water temperature and dissolved oxygen concentration at the seven lake sites, with the model fitted for all sites together.

Team members: Abigail S.L. Lewis, Caleb J. Robbins

Code repository:

https://github.com/eco4cast/Forecast_submissions/blob/main/Generate_forecasts

LM-precip (tg_precip_lm)

The tg_precip_lm model is a linear model fit using the function lm() in R. This is a very simple model with only total precipitation used as a model covariate. The model is fit every day as new data are made available, and is fit separately for each site/variable combination.

Driver uncertainty was included by using the 31 ensemble members from the NOAA GEFS weather forecast. No uncertainties from initial conditions or model process were included in the overall forecast uncertainty.

This model was used to forecast water temperature and dissolved oxygen concentration at the seven lake sites, with the model fitted separately for each site.

Team members: Abigail S.L. Lewis, Caleb J. Robbins

Code repository:

https://github.com/eco4cast/Forecast_submissions/blob/main/Generate_forecasts

LM-precip-all (tg_precip_lm_all_sites)

The tg_precip_lm_all_sites model is a linear model fit using the function lm() in R. This is a very simple model with only one covariate: total precipitation. The model is fit every day as new data are made available, and is fit across all sites, using site ID as a factor in the regression.

Driver uncertainty was included by using the 31 ensemble members from the NOAA GEFS weather forecast. No uncertainties from initial conditions or model process were included in the overall forecast uncertainty. This model was used to forecast water temperature and dissolved oxygen concentration at the seven lake sites, with the model fitted for all sites together.

Team members: Abigail S.L. Lewis, Caleb J. Robbins

Code repository:

https://github.com/eco4cast/Forecast_submissions/blob/main/Generate_forecasts

LM-temp (tg_temp_lm)

The tg_temp_lm model is a linear model fit using the function lm() in R. This is a very simple model with only one covariate: air temperature. The model is fit every day as new data are made available, and is fit separately for each site/variable combination.

Driver uncertainty was included by using the 31 ensemble members from the NOAA GEFS weather forecast. No uncertainties from initial conditions or model process were included in the overall forecast uncertainty.

This model was used to forecast water temperature and dissolved oxygen concentration in the seven lake sites, with the model fitted separately for each site.

Team members: Abigail S.L. Lewis, Caleb J Robbins

Code repository:

https://github.com/eco4cast/Forecast_submissions/blob/main/Generate_forecasts

LM-temp-all (tg_temp_lm_all_sites)

The tg_temp_lm_all_sites model is a linear model fit using the function lm() in R. This is a very simple model with only one covariate: air temperature. The model is fit every day as new data are made available, and is fit across all sites, using site ID as a factor in the regression.

Driver uncertainty was included by using the 31 ensemble members from the NOAA GEFS weather forecast. No uncertainties from initial conditions or model process were included in the overall forecast uncertainty.

This model was used to forecast water temperature and dissolved oxygen concentration at the seven lake sites, with the model fitted for all sites together.

Team members: Abigail S.L. Lewis, Caleb J. Robbins

Code repository:

https://github.com/eco4cast/Forecast_submissions/blob/main/Generate_forecasts

Random Forest (tg_randfor)

Random Forest is a machine learning model that is fitted with the ranger() function in the *ranger* R package (Wright & Ziegler 2017) within the *tidymodels* framework (Kuhn & Wickham 2020). The model drivers are unlagged air temperature, air pressure, relative humidity, surface downwelling longwave and shortwave radiation, precipitation, and northward and eastward wind. Only data prior to 2023-01-01 were used for any model training; similarly, model fits were not updated with any 2023 data when generating forecasts in 2023. Hyperparameters were selected for each site using 10-fold cross validation (repeated 5 times per site), selecting the hyperparameter combination with the lowest average RMSE. The number of trees was set to 500 but we tuned two hyperparameters for a) the minimum node

size for each tree and b) the number of randomly selected predictors. Model predictions are independent in time. The random forest model predicts observations for every NOAA GEFS ensemble member and forecast horizon of the predicted drivers, so only driver uncertainty is represented.

This model was used to forecast water temperature and dissolved oxygen concentration in the seven lake sites, with the model fitted separately for each site.

Team members: Abigail S.L. Lewis, Caleb J. Robbins

Code repository:

https://github.com/eco4cast/Forecast_submissions/blob/main/Generate_forecasts/tg_randfor/train_model.R

https://github.com/eco4cast/Forecast_submissions/blob/main/Generate_forecasts/tg_randfor/forecast_model.R

Lasso (tg_lasso)

Lasso is a machine learning model implemented in the same workflow as tg_randfor, but with different hyperparameter tuning. The model drivers are unlagged air temperature, air pressure, relative humidity, surface downwelling longwave and shortwave radiation, precipitation, and northward and eastward wind. Only data prior to 2023-01-01 were used for any model training; similarly, model fits were not updated with any 2023 data when generating forecasts in 2023. Hyperparameters were selected for each site using 10-fold cross validation (repeated 5 times per site), selecting the hyperparameter combination with the lowest average RMSE. Lasso regressions were fitted with the function glmnet() in the package *glmnet* (Tay et al. 2023), where the regularization hyperparameter (lambda) is tuned and selected with 10-fold cross validation.

This model was used to forecast water temperature and dissolved oxygen concentration at the seven lake sites, with the model fitted separately for each site.

Team members: Abigail S.L. Lewis, Caleb J. Robbins

Code repository:

https://github.com/eco4cast/Forecast_submissions/blob/main/Generate_forecasts/tg_lasso/train_model.R\
https://github.com/eco4cast/Forecast_submissions/blob/main/Generate_forecasts/tg_lasso/forecast_model.R

XGBoost (xgboost_parallel)

The XGBoost model is an extreme gradient boosted random forest (XGBoost) machine learning model that uses predicted atmospheric conditions and day of year as covariates. This model utilises the *xgboost* R package (Chen & Guestrin 2016; Chen et al., 2023).

A new model was trained for each site daily using air temperature, solar radiation (surface downwelling shortwave flux in air), relative humidity, and day of year. Models were trained on a random sample of 80% of the historic data, reserving 20% for evaluation. Models have 15 trees and for each tree to have a maximum depth of 10. The model was then evaluated on the remaining samples, the error variance being recorded. A forecast with 31 ensemble members was generated for each day in the forecasting horizon using the ensemble members from the NOAA GEFS weather forecast, representing driver uncertainty. Those predictions then have normally distributed random noise added to them matching the recorded error variance. Model uncertainty is derived from NOAA ensemble members as well as random noise based on estimated model accuracy (process uncertainty).

This model was used to forecast water temperature and dissolved oxygen concentration at the seven lake sites, with the model fitted separately for each site.

Team members: Gregory Harrison, R. Quinn Thomas

Code Repository:

Original: https://github.com/Grepath/XGBoostNeon4Casts/blob/main/Aquatics_ParallelXGB.R

Forked (running automation): https://github.com/FLARE-forecast/XGBoostNeon4Casts/blob/main/Aquatics_ParallelXGB.R

References

Blanchard, G. F., Guarini, J. M., Richard, P., Gros, P., & Mornet, F. (1996). Quantifying the short-term temperature effect on light- saturated photosynthesis of intertidal microphytobenthos. Marine Ecology Progress Series, 134(1–3), 309–313. https://doi.org/10.3354/meps134309

Chapra, S.C. (2008). Surface Water-Quality Modeling. Waveland Press, Inc.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. (2023). xgboost: Extreme Gradient Boosting. R Package Version 1.7.5.1. Retrieved from https://cran.r-project.org/package=xgboost

Eppley, R. W. (1972). Temperature and phytoplankton growth in the sea. Fishery Bulletin, 70(4), 1063–1085.

Evensen, G. (2003). The Ensemble Kalman Filter: theoretical formulation and practical implementation. Ocean Dynamics, 53(4), 343–367. https://doi.org/10.1007/s10236-003-0036-9

Goudsmit, G.H., Burchard, H., Peeters, F., & Wüest, A. (2002). Application of k-ε turbulence models to enclosed basins: The role of internal seiches. Journal of Geophysical Research, 107(C12), 23-1–23-13. https://doi.org/10.1029/2001JC000954

Hamill, T. M., Whitaker, J. S., Shlyaeva, A., Bates, G., Fredrick, S., Pegion, P., et al. (2022). The Reanalysis for the Global Ensemble Forecast System, Version 12. Monthly Weather Review, 150(1), 59–79. https://doi.org/10.1175/MWR-D-21-0023.1

Herzen, J., Lässig, F., Piazzetta, S. G., Neuer, T., Tafti, L., Raille, G., et al. (2022). Darts: User-Friendly Modern Machine Learning for Time Series. Journal of Machine Learning Research, 23(124), 1–6. Retrieved from http://jmlr.org/papers/v23/21-1177.html

Hinshelwood C. N. (1945) The chemical kinetics of bacterial cells. Clarendon Press, Oxford

Hipsey, M. R., Bruce, L. C., Boon, C., Busch, B., Carey, C. C., Hamilton, D. P., et al. (2019). A General Lake Model (GLM 3.0) for linking with high-frequency sensor data from the Global Lake Ecological Observatory Network (GLEON). Geoscientific Model Development, 12(1), 473–523. https://doi.org/10.5194/gmd-12-473-2019

Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software, 27(3). https://doi.org/10.18637/jss.v027.i03

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hare-Wild, M., et al. (2023). forecast: Forecasting functions for time series and linear models. R package version 8.21.1. Retrieved from https://pkg.robjhyndman.com/forecast/

Kuhn M & Wickham H (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles, https://www.tidymodels.org

Livingstone, D. M., & Imboden, D. M. (1989). Annual heat balance and equilibrium temperature of Lake Aegeri, Switzerland. Aquatic Sciences, 51(4), 351–369. https://doi.org/10.1007/BF00877177

Martin, J.L. and McCutcheon, S. (1998). Hydrodynamics and Transport for Water Quality Modeling. CRC Press

Monod, J. 1950. "Technique, Theory and Applications of Continuous Culture." Annales de l'Institut Pasteur 79 (4): 390–410.

Moore, T. N., Mesman, J. P., Ladwig, R., Feldbauer, J., Olsson, F., Pilla, R. M., et al. (2021). LakeEnsemblR: An R package that facilitates ensemble modelling of lakes. Environmental Modelling & Software, 143, 105101. https://doi.org/10.1016/j.envsoft.2021.105101

Moore, T. N., Mesman, J. P., Ladwig, R., Feldbauer, J., Olsson, F., Pilla, R. M., et al. (2021). LakeEnsemblR: An R package that facilitates ensemble modelling of lakes. Environmental Modelling & Software, 143, 105101. https://doi.org/10.1016/j.envsoft.2021.105101

Moulton TL (2018). rMR: Importing Data from Loligo Systems Software, Calculating Metabolic Rates and Critical Tensions. R package version 1.1.0, https://CRAN.R-project.org/package=rMR

Norberg, J. (2004). Biodiversity and ecosystem functioning: A complex adaptive systems approach. Limnology and Oceanography, 49, 1269–1277. https://doi.org/10.4319/lo.2004.49.4_part_2.1269

O'Hara-Wild, M., Hyndman, R., & Wang, E. (2022). fable: Forecasting Models for Tidy Time Series. R package version 0.3.2. Retrieved from https://cran.r-project.org/package=fable

Rosso, L., Lobry, J. R., & Flandrois, J. P. (1993). An Unexpected Correlation between Cardinal Temperatures of Microbial Growth Highlighted by a New Model. Journal of Theoretical Biology, 162(4), 447–463. https://doi.org/10.1006/jtbi.1993.1099

Steele, J. H. (1962). Environmental Control of Photosynthesis in the Sea. Limnology and Oceanography, 7(2), 137–150. https://doi.org/10.4319/lo.1962.7.2.0137

Tay, J. K., Narasimhan, B., & Hastie, T. (2023). Elastic Net Regularization Paths for All Generalized Linear Models. Journal of Statistical Software, 106(1). https://doi.org/10.18637/jss.v106.i01

Taylor, S. J., & Letham, B. (2018). Forecasting at Scale. The American Statistician, 72(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

Thomas, R. Q., Figueiredo, R. J., Daneshmand, V., Bookout, B. J., Puckett, L. K., & Carey, C. C. (2020). A Near-Term Iterative Forecasting System Successfully Predicts Reservoir Hydrodynamics and Partitions Uncertainty in Real Time. Water Resources Research, 56, e2019WR026138. https://doi.org/10.1029/2019WR026138

Thomas, R. Q., McClure, R. P., Moore, T. N., Woelmer, W. M., Boettiger, C., Figueiredo, R. J., et al. (2023). Near-term forecasts of NEON lakes reveal gradients of environmental predictability across the US. Frontiers in Ecology and the Environment, 21(5), 220–226. https://doi.org/10.1002/fee.2623

Umlauf, L., Burchard, H., & Bolding, K. (2005). GOTM - Sourcecode and Test Case Documentation. Retrieved from http://www.gotm.net/pages/documentation/manual/stable/pdf/a4.pdf

Verburg, P., & Antenucci, J. P. (2010). Persistent unstable atmospheric boundary layer enhances sensible and latent heat loss in a tropical great lake: Lake Tanganyika. Journal of Geophysical Research, 115(D11), D11109. https://doi.org/10.1029/2009JD012839

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1). https://doi.org/10.18637/jss.v077.i01